DOCUMENT RESUME

ED 367 152 FL 021 874

AUTHOR Nunan, David
TITLE Second Language Proficiency Assessment and Program Evaluation.
PUB DATE 91
NOTE 15p.; In: Anivan, Sarinee, Ed. Issues in Language Programme Evaluation in the 1990's. Anthology Series 27; see FL 021 869.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Criterion Referenced Tests; Educational Objectives; Evaluation Criteria; Foreign Countries; *Language Proficiency; *Language Tests; *Program Evaluation; Second Language Instruction; *Second Language Programs; Student Evaluation

ABSTRACT

A discussion of the role of second language proficiency assessment in the evaluation of language programs argues that for four reasons, the use of proficiency is inappropriate as a central element in evaluation. The reasons are: (1) the construct of proficiency has not been operationalized in a way that enables it to be used usefully; (2) criterion-referenced measures of achievement are of more practical utility than are statements of proficiency not tied to specific program goals; (3) regardless of the terms in which learner outcomes are to be defined, comprehensive program evaluation requires the collection, interpretation, and evaluation of data relating to a range of processes and elements operating within a particular educational context, not just learner outcomes; and (4) process data is needed to interpret outcomes data. A number of practical suggestions for program evaluation, and sample instruments, are offered. (MSE)

SECOND LANGUAGE PROFICIENCY ASSESSMENT AND PROGRAM EVALUATION

David Nunan

INTRODUCTION

I have been asked, today, to examine the role of second language proficiency assessment in program evaluation. In the paper, I shall argue that while assessment is an important component of program evaluation, it is only one component. I shall further argue against the construct of general 'curriculum-free' proficiency, as this is currently operationalized in the literature, as a central component in program evaluation. 'Curriculum-free' proficiency is proficiency which is not tied to or referenced against curriculum goals. My reservations about the use of 'proficiency', thus conceived, as a central element in program evaluation are four in number, and will be expanded upon in the course of the presentation.

1 The construct of proficiency has not been operationalized in a way which enables it to be usefully used for the purposes of program evaluation.

2 Criterion-referenced measures of achievement are of more practical utility than statements of proficiency which are not related to program goals.

3 Regardless of the terms in which learner outcomes are to be defined, comprehensive program evaluation requires the collection, interpretation and evaluation of data relating to a range of processes and elements operating within a particular educational context, not just learner outcomes.

4 In order to interpret outcome data, one needs process data.

The paper contains a number of practical suggestions which have implications for carrying out program evaluation within a Southeast Asian context, and includes some sample instruments for carrying out such evaluations.

THE CONCEPTS OF LANGUAGE PROFICIENCY AND EVALUATION

This paper is centrally concerned with proficiency assessment and evaluation, and I should therefore attempt to clarify my understanding of these terms from the outset. In some educational systems, the terms 'assessment' and 'evaluation' are used interchangeably - witness the following quote from Gronlund:

> Evaluation may be defined as a systematic process of determining the extent to which instructional objectives are achieved by pupils. There are two important aspects of this definition. First, note that evaluation implies a systematic process, which omits casual, uncontrolled observation of pupils. Second, evaluation assumes that instructional objectives have been previously identified. Without previously determined objectives, it is difficult to judge clearly the nature and extent of pupil learning.
>
> (Gronlund 1981:5)

Gronlund, in circumscribing evaluation in terms of learning outcomes, presents an extremely narrow input-output view of evaluation and, by extension, education. In fact, he is using the term 'evaluation' roughly in the sense in which I would use 'assessment'. I would like to suggest that, while they are obviously related, they mean rather differ · · things. To me, assessment refers to the set of processes through which we make judgements about what a learner is able to do in the target language. We may or may not assume that such abilities have been brought about by a program of study.

'Evaluation' is a wider term than 'assessment'. While it entails the collection of information on what learners can do in the target language it also involves additional processes designed to assist us in interpreting and acting on the results of our assessment. The data resulting from evaluation assist us in deciding whether a course needs to be modified or altered in any way so that objectives can be achieved more effectively. If certain learners are not achieving the goals and objectives set for a course, it is necessary to determine why this is so. We would also wish, as a result of evaluating a course, to have some idea about what measures might be taken to remedy any shortcomings. Evaluation, then, is not simply a process of obtaining information, it is also a decision-making process.

In this area, there seems to be a certain tension between 'measurement' and 'evaluation'. Those who are seduced by the illusion of certainty offered by tools and techniques for measuring things sometimes seem to forget that there is an essential difference between the value neutral processes of measurement and the value laden nature of evaluation (Wolf 1984).

Thus far, I have argued that assessment is a process of collecting information about what a learner can do in the target language, while program evaluation is a more general process of obtaining a variety of information relating to different curriculum elements and processes, for decision-making purposes. For most evaluations, I believe it is useful to collect data on what learners can and cannot do, although this view is by no means universally held by program evaluators, and for some types of evaluation it may be either unnecessary or impossible to obtain such data.

In recent years, a great deal has been written and said about the use of measures of proficiency as a means of assessing learners. I believe that there are some serious problems with the way the concept of proficiency has been defined and operationalised, and in this section I shall explore some of these problems. This will provide a basis for considering the feasibility or desirability of adopting a 'program-free'approach to proficiency assessment. Before we consider assessment instruments themselves, however, it is necessary to engage in some terminological ground clearing.

Within the literature, there is considerable confusion about the constructs and terminology associated with language development and use. Confusion, disagreement and uncertainty are reflected in much of the writing associated with language testing, a confusion which can be partly explained by a lack of agreement about the nature of language, language learning and use. This confusion is evident in the various ways in which terms such as 'competence', 'performance', 'proficiency' and so on are used. Although he did not create the terms, Chomsky (1965) gave prominence to the notions of 'competence' and 'performance'. For Chomsky, 'competence' refers to the mastery of principles governing language bahaviour. 'Performance, on the other hand, refers to the manifestation of these internalised rules in actual language use. The terms have come to be used to refer to what a person knows about a language (competence) in contrast to what that person does (performance). More recently the term 'communicative competence' has gained currency, and there has been some debate as to the actual constituents of this construct. There is also considerable ongoing debate about what it means to 'know the rules of a given language'.

Diller (1978) attempts to resolve this paradox by suggesting that knowledge exists on a subconscious level:

> ... if children are not able to formulate the rules of grammar which they use, in what sense can we say that they 'know' these rules? This is the question which has bothered linguists. The answer is that they know the rules in a functional way, in a way which relates the changes in abstract grammatical structure to changes in meaning. Knowledge does not always have to be consciously formulated. Children can use tools before they learn the names for these tools.
>
> (Diller 1978: 26-27)

If we accept that knowledge need not be consciously formulated, but may manifest itself in the ability to use the language, it would seem to render the competence-performance distinction rather uncertain. (See also the systemic-functionalist view that the distinction is unnecessary and misleading because language is what language does.)

Krashen (1981, 1982) further confuses the issue by suggesting that knowledge of linguistic rules is the outward manifestation of one psychological construct (learning), while use of these rules to communicate is the manifestation of another construct (acquisition).

Rea (1985) subsequently questioned the need for a 'competence' construct by suggesting that as we can only observe instances of performance, not competence, the competence-performance distinction is redundant. In testing terms, she suggests that we forget about 'competence' and think in terms of communicative performance and non-communicative performance.

This brings us to the point where linguistic knowledge is to be defined in terms of what an individual is able to do with that knowledge. This is reflected in the competency-based ESL movement which has gained a certain amount of prominence, particularly in the United States. As though there were not enough confusion over terminology, this movement is using 'competence' to refer to things learners can do with language; that is, it is used in roughly the same sense as 'performance' in the earlier competence-performance distinction. In ESL, 'a competency is a task-oriented goal written in terms of behavioural objectives' (CAL 1983:9) which has clear implications for assessment. Assessment is built in. Once the competency has been identified, it also serves as a means of evaluating student performance. Since it is performance based, assesment rests on whether the student can perform the competency or not. The only problem is to establish the level at which the student can perform the competency. (op cit:11-13)

Within the literature, some writers use the term 'proficiency' as an alternative to 'competency' (see, for example Higgs 1984). Richards, however, makes a clear distinction between 'competence' and 'proficiency', although he characterises the concept of proficiency in the same way as Competency Based Education characterises competency:

> 1 When we speak of proficiency, we are not referring to knowledge of a language, that is, to abstract, mental and unobservable abilities. We are referring to performance, or, that is, to observable or measurable behaviour. Whereas competence refers to what we know about the rules of use and rules of speaking of a language, proficiency refers to how well we can use such rules in communication.
>
> 2 Proficiency is always described in terms of real-world tasks, being defined with reference to specific situations settings purposes activities and so on.
>
> (Richards 1985: 5)

Richards goes on to argue that:

> A proficiency-oriented language curriculum is not one which sets out to teach learners linguistic or communicative competence, since these are merely abstractions or idealisations: rather, it is organised around the particular kinds of communicative tasks the learners need to master and the skills and bahaviours needed to accomplish them. The goal of a proficiency-based curriculum is not to provide opportunities for the learners to 'acquire' the target language: it is to enable learners to develop the skills needed to use language for specific purposes.
>
> (Richards 1985: 5)

In this section, I have attempted to highlight some of the confusion surrounding key concepts relating to the nature of language proficiency. This confusion is due partly to the inconsistent application of terms to concepts and partly to confusion over the nature of the concepts themselves. If we follow the portrayal of Richards, proficiency, simply put, refers to the ability to perform real-world tasks with a prespecified degree of skill. In programmatic terms this definition is probably reasonable enough. However, when it comes to the assessment of second language proficiency, the psychological reality of the construct become problematic, as we shall now see.

In order to assess any area of human behaviour, it is necessary to have some idea of what it is we are trying to assess. What is it that testers of language proficiency are trying to assess? We can get some idea by looking at the instruments which have been developed. One increasingly popular instrument is the proficiency rating scale. What follows is the generic description of speaking proficiency at an intermediate-high level. It is taken from the American Council on the Teaching of Foreign Languages Provisional Proficiency Guidelines.

Able to satisfy most survival needs and limited social demands.
Shows some spontaneity in language production but fluency is very uneven.
Can initiate and sustain a general conversation but has little understanding of the social conventions of conversation.
Developing flexibility in a range of circumstances beyond immediate survival needs.
Limited vocabulary range necessitates much hesitation and circumlocution.
The commoner tense form occur but are frequent in formation and selection.
Can use most question forms.
While some word order is established, errors still occur in more complex patterns.
Cannot sustain coherent structures in longer utterances or unfamiliar situations.
Ability to describe and give precise information is limited.
Aware of basic cohesive features such as pronouns and verb inflections, but many are unreliable, especially if less immediate in reference.
Extended discourse is largely a series of short, discrete utterances.
Articulation is comprehensible to native speakers used to dealing with foreigners, and can combine most phonemes with reasonable comprehensibility, but still has difficulty in producing certain sounds in certain positions or in certain combinations, and speech will usually be laboured.
Still has to repeat utterances frequently to be understood by the general public.
Able to produce some narration in either past or future.
(Cited in Savignon and Berns 1984: 228-229)

The use of such scales is fraught with hidden dangers, which, for reasons of space, can only be briefly sketched out here. The scales themselves tend to take on ontological status - that is, there is a tendency to assume that such a person as an 'Intermediate-High' learner actually exists and that there is such a thing as 'Intermediate-High' ability - rather than being something constructed to account for observable or hypothetical features of learners' speech. (See also, Lantolf and Frawley, 1988 who point out the essential circularity of the descriptions). The scales themselves have not always been empirically validated to determine if learners really do act in the ways described by the scales. Research from second language language acquisition is often overlooked or ignored. (Some scales actually violate findings from SLA research.) One rating scale (the Australian Second Language Proficiency Rating Scale) makes claims about the equivalence of real world tasks and their appropriacy at different levels of proficiency. It is suggested, for example, that the tasks of 'returning an unsatisfactory purchase' and 'explaining some personal symptoms to a doctor' are of the same order of difficulty. However, no empirical evidence is provided that these tasks draw on the same linguistic and communicative resources, nor that the ability to perform such tasks can be determined by indirect measures of proficiency such as an oral interview. Finally, in terms of construct validity, the scales confound phonological, morphosyntactic, lexical, semantic and pragmatic features.

Program-free proficiency assessment and learner achievement

Within the literature, there are claims that program evaluation should be based on tests of general language proficiency through means such as the proficiency rating scales critiqued in the preceding section, not on achievement measures which are related to or associated with the program being evaluated. This line of argument is based on the view that unless transfer of learning can be demonstrated to have taken place, then learning, in any meaningful sense can not be said to have taken place. ('Transfer' is generally defined as the extent to which knowledge and skills developed in one field can be taught in a way which enables them to be utilized in another field.) There are a number of problems associated with the above argument, as we shall shortly see. In fact, even if learning transfer can be demonstrated to have occurred, it is quite another matter to demonstrate that learning is the result of a specific program intervention.

The whole issue of transfer of learning has, of course, been long debated in the educational and cognitive psychology literature. One debate concerns the relative claims of the cognitive skills transfer hypothesis versus the subject-domain hypothesis. The cognitive skills transfer hypothesis suggests that the development of knowledge and skills in certain subject domains can develop general learning and thinking skills which will transfer to other subject domains. For example, in a Western context, the teaching of languages, particularly Latin and Greek, was, for many years, defended on the grounds that it facilitated the

development of reasoning skills which could be subsequently employed on more relevant subject areas. However, there has never been any evidence to support this claim. In fact, what evidence there is seems to run counter to the claim (see, for example, Thorndike and Woodward 1981, and Resnick 1987 cited in De Corte 1987). In contrast to the paucity of data on the transferability of general learning skills, there is a great deal of evidence to suggest that "the availability and flexible use of a well-ordered body of domain-specific knowledge play a major role in successful learning and problem-solving activities." (Glaser 1987).

Voss (1987) provides a reconceptualisation of the concepts of learning and transfer based upon a general information processing model of problem solving which suggests that learning and acquisition are subordinate to transfer. His paper begins with an analysis of the concepts 'acquisition', 'learning' and 'transfer', as defined by Association Theory which derived its definitions from everyday knowledge rather than systematic analysis. 'Acquisition' was investigated in "multiple trila experiments which intrinsically presumed contiguity and frequency as the mechanisms producing acquisition". 'Learning' was defined as an improvement in performance as a result of practice, while 'transfer' was defined as "the influence of the learning of one task upon the performance of a second task" (Voss 1987: 608). With the demise of associationism came a decrease in the use of multiple trial acquisition experiments and the use of the concepts 'learning', 'retention' and 'transfer'.

Voss outlines Jenkins' tetrahedal model which suggests that learning and memory are dependent on the interaction between four classes of variables. These are 'orienting task' (e.g. instructions, activities); materials (e.g. sensory mode, physical structure); criterial tasks (e.g. recall, recognition, problem-solving); subject characteristics (e.g. activities, interests, knowledge). As the manipulation of two or more of these variables results in a significant interaction, it is almost impossible to conduct laboratory experiments which will yield generalisable results. The thrust of Jenkins' work is to suggest that:

> ... there is no one way to learn since learning wil depend on the instructional task, the materials, the criterion of learning and the characteristics of the individual who is learning. The answer to the question of how best to teach a particular subject matter to a particular group of subjects becomes "it depends".
>
> (Voss 1987: 609)

Given these criticisms, Voss sets out to reconceptualise the key concepts of learning, retention and transfer. He adopts a phenomenological stance, suggesting that individual differences such as intelligence, prior knowledge and experience, attitudes and cognitive skills will have a crucial effect on what is learned and retained. The reason why true experiments come up with few substantive findings is that they employ procedures to randomise the very individual differences which determine what is learned and what is not. Beretta (1986) has made similar points in his call for the use of field rather than laboratory experimentation in language program evaluation.

Returning to the domain of language, rather than the more broadly conceived cognitive domain, the argument for program-free assessment is, to my mind rather curious. If the purpose of providing learners with a language education is to enable them to carry out a range of communicative tasks in that language, then it would seem entirely proper to base one's assessment on the achievement of specific curricular goals rather than on vaguely formulated notions of proficiency operationalised through proficiency scales and other tests of dubious validity. Such a suggestion is consonant with current trends in assessment outlined by Baumgart (1987):

> - a concerted move towards some form of standards-based assessment;
> - a growth in school-level initiatives in assessment and rep rting, including quite widespread use of profiles, records of achievement and goal-based asessment;
> - much closer links between curricula and assessment with an emphasis on formative assessment;
> an emphasis on positive achievement and attempts to negotiate tasks and objectives which stretch students' capabilities but which also offer a reasonable chance of success;
> - consideration of the use of summative system-level records, albeit produced by schools, to underwrite and supplement formal certificates.
>
> (Cited in Brindley 1989: 93)

Brindley (1989) provides an invaluable source book of practical ideas, suggestions and illustrations of ways of incorporating criterion-related assessment instruments into the curriculum. He provides samples of performance profiles, records of achievement, graded objectives, rating scales, self-assessment checklists. Examples of such instruments from Nunan (1988) and Scarino et al. (1988) are provided in an appendix to the paper. Brindley himself has written extensively on the distinction between achievement testing and proficiency testing, arguing that the division fails to capture the range of purposes for which assessment may be carried out, and, further, that it fails to distinguish between the type and level of information. He attempts to resolve the tension between the two concepts by postulating three different types of achievement / proficiency. Of these, only the first is "program-free". (Clark has coined the term "prochievement" to capture the idea of ongoing communicative assessment that is related to the program's proficiency goals.

> Level 1: Achievement of overall proficiency in a particular language skill or skills ("general" proficiency)
> Level 2: Achievement of particular proficiency-related objectives as part of a given course ("functional")
> Level 3: Achievement of specific objectives relating to knowledge and enabling skills taught in a particular course ("structural") (Brindley 1989).

Thus far, I have analysed and critiqued the notion of utilizing curriculum-free proficiency measures as means of assessing student progress. I have outlined some of the conceptual problems of the concept itself, as well as pointing out some of the inadequacies of instruments for measuring general language proficiency. It should be clear, therefore that I do not accept the validity of using such measures for the purposes of program evaluation. I would also refer you to Bachman's discussion on objectives-based and program-free evaluation. In the rest of the paper, I should like to focus more directly on program evaluation, and suggest that, while the incorporation of criterion-referenced assessment measures should form part of any adequate evaluation process, that they should not form the whole, or even the major part of the evaluation process. The two principal justifications I should like to offer for this assertion are (1) that evaluation involves much more than simply monitoring and measuring learning progress, and (2) that evaluation needs to focus on instructional processes as much as learning outcomes.

In concluding this section, I should like to point out that the use of individual gain scores to determine program effectiveness is not only problematic on theoretical grounds, but also on the practical grounds that gain scores are often not picked up due to the grossness of the measureing instruments. Within the Australian Adult Migrant Education Program, there are instances in which proficiency scores are actually lower at the end of a course than at the beginning!

The scope of program evaluation

In this paper, I have argued against a narrow input-output view of program evaluation, which references evaluation solely against learner output. The breadth and scope of any program evalaution must be referenced against two two important questions: "Who wants to know?" and "Why do they want to know?" As Cronbach has said, in his call for a reformulation and transformation in evaluation:

> The proper mission of evaluation is not to eliminate the fallibility of authority or to bolster its credibility. Rather, its mission is to facilitate a democratic, pluralistic process by enlightening all the participants. ... The evaluator is an educator; his success is to be judged by what others learn. .... Scientific quality is not the principal standard; an evaluation should aim to be comprehensible, correct and complete, and credible to partisans on all sides.
>
> (Cronbach 1980: 1, 11)

Assuming that most evaluations are not simply tokenistic exercises in indictment or exoneration, then program evaluators will want not only / even 'proof' in product terms, but 'insights' into the curicular processes and dynamics giving rise to particular outputs. In order to generate such insights, questions needs to be asked, and data gathered, on different aspects of the curriculum. Any area of the curriculum can be evaluated, from initial program

planning through to the assessment / evaluation processes themselves. Some of the questions which might be posed in relation to different curriculum areas are set out in Table 1, which has been extracted from Nunan 1988.

Table 1
Some key questions in program evaluation

Curriculum area	Sample Questions
The Planning Process Needs Analysis	Are the needs analysis procedures effective? Do they provide useful information for course planning? Do they provide useful data on subjective and objective needs? Can the data be translated into content?
Content	Are goals and objectives derived from needs analysis? If not, from where are they derived? Are they appropriate for the specified groups of learners? Do the learners think the content is appropriate? Is the content appropriately graded? Does it take speech processing constraints into account?
Implementation Methodology	Are the materials, methods and activities consonant with the prespecified objectives? Do the learners think the materials, methods and activities are appropriate?
Resources	Are resources adequate / appropriate?
Teacher	Are the teacher's classroom management skills adequate?
Learners	Are the learning strategies of the students efficient? Do learners attend regularly? Do learners pay attention / apply themselves in class? Do learners practise their skills outside the classroom? Do the learners appear to be enjoying the course? Is the timing of the class and the type of learning arrangement suitable for the students?

	Do learners have personal problems which interfere with their learning?
Assessment and evaluation	Are the assessment procedures appropriate to the prespecified objectives? Are there opportunities for self-assessment by learners? If so, what? Are there opportunities for learners to evaluate aspects of the course such as learning materials, methodology, learning arrangement? Are there opportunities for self-evaluation by the teacher?

As I have already pointed out, in any evaluation, estimating the extent of learning outcomes is only a first step. Working out why certain learners have not achieved program goals is a much more difficult process requiring interpretation and analysis. In a study into teacher perceptions of the causes of learner failure reported in Nunan (1988), a group of ESL teachers were asked to nominate those causes which they felt were significant factors in the failure of learners to achieve program goals. The results of this investigation are summarised in Table 2. I have subcategorised these into causes attributable to the learner and causes attributable to the teacher.

Table 2

Survey results of causes of learner failure (After Nunan 1988)

Cause	Percentage of teachers rating this as a significant factor in learner failure
Causes attributable to the learner	
Inefficient learning strategies	77
Failure to use language out of class	77
Irregular attendance	45
Particular macroskill problems	32
Poor attention in class	9
Personal (non-language) problems	9
Learner attitude	4
Causes attributable to the teacher	
Inappropriate learning activities	32
Inappropriate objectives	27
Faulty teaching	23

From the data, it can be seen that, in general, the teachers surveyed saw responsibility for failure residing largely with the learners. (Although it is worth noting that, in relation to causes attributable to the teacher, one third of those surveyed identified inappropriate learning activities as a possible cause, and approximately a quarter identified inappropriate objectives and faulty teaching as having a significant effect on learning outcomes.)

The Need for Process Data in Program Evaluation

In order to validate the sorts of observations yielded by the study reported above, it is important to obtain data about learning and teaching processes themselves. Systematic observation is one important means of collecting such data. Non-observable problems such as failure to activate language out of class can be collected through learner diaries and self-reports. Other techniques, which are described and illustrated in some detail in Nunan (1989) include interviews and questionnaires, protocol analysis, transcript analysis, stimulated recall, and seating chart observation records. Ideally, a number of such techniques and instruments should be utilized in order to obtain multiple perspectives on the program under investigation.

The desirability of obtaining data on program outcomes and teaching processes is illustrated in a study reported in Spada (1990). This investigation sought to determine (a) how different teachers interpreted theories of communicative language teaching in terms of their classroom practice, and (b) whether different classroom practices had any effect on learning outcomes. Three teachers and their intermediate "communicatively-based" ESL classes were used in the study. Each class was observed for five hours a day, once a week, over a six-week period. Students were given a battery of pre- and post-tests including the Comprehensive English Language Test and the Michigan Test of English Language Proficiency. The study utilized the COLT observation scheme as well as a qualitative analysis of classroom activity types. This indicated that one of the classes, Class A, differed from the other two in a number of ways:

> A spent considerably more time on form-based activities (with explicit focus on grammar), while classes B and C spent more time on meaning-based activities (with focus on topics other than language). Classes B and C also had many more authentic activity types than class A. Furthermore, the classes differed in the way in which certain activities were carried out, particularly listening activities. For example, in classes B and C, the instructors tended to start each activity with a set of predictive exercises. These were usually followed by the teacher reading comprehension questions to prepare the students for the questions they were expected to listen for. The next step usually involved playing a tape-recorded passage and stopping the tape when necessary for clarification and repetition requests. In class A, however, the listening activities usually proceeded by giving students a list of comprehension questions to read silently; they could ask teachers for assistance if they had difficulty understanding any of them. A tape-recorded passage was then played in its entirety while students answered comprehension questions.
>
> (Spada 1990: 301)

The qualitative analysis confirmed the class differences, showing, for example, that class A spent twice as much time on form-based work than class C and triple the time spent by class B. To investigate whether these differences contributed differently to the learners L2 proficiency, pre- and post-treatment test scores were compared in an analysis of covariance. Among other things, results indicated that groups B and C improved their listening significantly more than group A, despite the fact that class A spent considerably more time in listening practice than the other classes.

Research such as that carried out by Spada indicated that there are in fact measurable differences in the way in which instruction is delivered in language programs which have similar ideological underpinnings, and that these differences can be related to learning outcomes. On a methodological level, it indicates that we need qualitative data based on classroom observation if we are to interpret, for the evaluative purposes of making decisions about program alternatives, the quantitative data yielded by assessment instruments of various sorts.

CONCLUSION

In this paper, I have taken a critical look at the role of second language proficiency assessment in program evaluation. I have examined some of the problematic aspects of the construct 'general language proficiency', as well as the theoretical and practical problems associated with attempting to measure such a construct. While I have referenced most of my

comments against rating scales of one type or another, they are also pertinent to other types of proficiency test. As an alternative, I have suggested that curriculum-bound, criterion-referenced forms of assessment be developed. Sample assessment instruments are appended to the paper.

Given the length, purpose and nature of this paper, it has not been possible to comment on the problems associated with criterion-referenced assessment. I refer you to the paper given at this conference by Brindley who addresses some of the problems of trying to ensure validity and reliability. For example, how many times must a learner be observed to be able to do something, under what conditions, with what constraints, and in what contexts?

Assesment is an important component of program evaluation. However, determining what learners have or have not gained from a program is only one aspect of the evaluation process. In the paper, we have seen some of the other curricular elements which may fruitfully form the subject of any comprehensive evaluation.

In the final part of the paper, I argued that we need to collect information on teaching processes as well as learning outcomes. Techniques for collecting such data are outlined, and a study illustrating the importance of having both process and product data is reported. Ultimately, the type of evidence which is collected, and the ways in which it is interpreted and reported must proceed with reference to the purpose, scope and nature of the evaluation itself. If the principal purpose is to provide data to funding authorities for accountability purposes, the processes and outcomes are likely to be significantly different from an evaluation designed to provide feedback to teachers or one aimed at the development of new materials and teaching techniques.

REFERENCES

Bachman, L. 1989. The development and use of criterion-referenced tests of language ability in language program evaluation. In R. K. Johnson (ed.) The Second Language Curriculum. Cambridge: Cambridge University Press.

Baumgart, N. 1987. Emerging trends. In Reports and Records of Achievement for School Leavers. Project Newsletter No. 2, April 1987.

Beretta, A. 1986. A case for field-experimentation in program evaluation. Language Learning, 36, 3

Brindley, G. 1989. Assessing Achievement in a Learner-Centred Curriculum. Sydney: National Centre for English Language Teaching and Research. CAL 1983. From the Classroom to the Workplace: Teaching ESL to Adults. Washington: Center for Applied Linguistics.

Chomsky, N. 1965. Aspects of the Theory of Syntax. Cambridge Mass.: M.I.T. Press.

Cronbach, L. 1980. Toward Reform of Program Evaluation. San Francisco: Josey-Bass.

De Corte, E. (ed.) 1987. Acquisition and transfer of knowledge and cognitive skills. International Journal of Educational Research, 11, 6.

De Cotre, E., H. Lodewijks, R. Parmentier and P Span (eds.) Learning and instruction. European research in an international context. Studia Pedagogica, 1.

Diller, K. 1978. The Language Teaching Controversy. Rowley Mass.: Newbury House.

Glaser, R. 1987. Learning theory and theories of knowledge. In E. De Corte et al. (eds.)

Gronlund, N. 1981. Measurement and Evaluation in Teaching. New York: Macmillan.

Higgs, T.V. (ed.) 1984. Planning for Proficiency: The Organising Principle. Lincolnwood: National Textbook Company.

Lantolf, J.P. and W. Frawley. 1988. Proficiency: understanding the construct. Studies in Second Language Acquisition, 10, 2.

Krashen S. 1981. Second Language Acquisition and Second Language Learning. Oxford: Pergamon.

Krashen, S. 1982. Principles and Practice is Second Language Acquisition. Oxford: Pergamon.

Nunan, D: 1988. The Learner-Centred Curriculum. Cambridge: Cambridge University Press.

Nunan, D. 1989. Understanding Language Classrooms. London: Prentice Hall.

Rea, P. 1985. Language Testing and the Communicative Language Teaching Curriculum. In

Y.P. Lee, C.Y.Y. Fook, R. Lord, and G. Low (eds.) New Directions in Language Testing. Oxford: Pergamon.

Resnick, L. 1987. Instruction and the cultivation of thinking. In E. De Corte et al. (eds.)

Richards, J. 1985. Planning for proficiency. Prospect, 1, 2.

Richards, J. C. and D. Nunan. (eds.) 1990. Second Language Teacher Education. Cambridge: Cambridge University Press.

Savignon, S. and M. Berns. (eds.) 1984. Initiatives in Communicative Language Teaching. Reading Mass.: Addison-Wesley.

Scarino, A., D. Vale, P. McKay and J. Clark. 1988. Evaluation, Curriculum Renewal and Teacher Development. Australian Language Levels Guidelines Book 4. Canberra: Curriculum Development Centre.

Spada, N. 1990. Observing classroom behaviours and learning outcomes. In J. C. Richards and D. Nunan. (eds.) 1990. Second Language Teacher Education. Cambridge: Cambridge University Press.

Thorndike, E. and R. Woodworth. 1981. The influence of improvement in one mental function upon the efficiency of other functions. Psychological Review, 8.

Voss, J. 1987. Learning and transfer in subject matter learning: A problem-solving model. International Journal of Educational Research, 11, 6.

Wolf, R. M. 1984. Evaluation in Education. New York: Praeger.

APPENDIX: Sample Criterion-Referenced Assessment Instruments

(Source: D. Nunan. 1988. The Learner-Centred Curriculum. Cambridge: Cambridge University Press.)

TABLE 9.1

Sample rating scales

Indicate the degree to which learners contribute to small-group discussions or conversation classes by circling the appropriate number.

(Key: 5 – outstanding, 4 – above average, 3 – average, 2 – below average, 1 – unsatisfactory)

1	The learner participates in discussions.	1	2	3	4	5
2	The learner uses appropriate non-verbal signals.	1	2	3	4	5
3	The learner's contributions are relevant.	1	2	3	4	5
4	The learner is able to negotiate meaning.	1	2	3	4	5
5	The learner is able to convey factual information.	1	2	3	4	5
6	The learner can give personal opinions.	1	2	3	4	5
7	The learner can invite contributions from others.	1	2	3	4	5
8	The learner can agree/disagree appropriately.	1	2	3	4	5
9	The learner can change the topic appropriately.	1	2	3	4	5

Rate the learner's speaking ability by circling the appropriate number.

1 2 3 4 5 6 7 8 9 10

Incapable of carrying out simple conversation

Carries out simple conversation giving personal information

Rate the learner's listening ability by circling the appropriate number.

1 2 3 4 5 6 7 8 9 10

Incapable of following simple instructions

Follows simple instructions in classroom setting

Checklist of reading skills

YES	NO	Recognises Roman script upper/lower case
YES	NO	Identifies numbers in various formats
YES	NO	Comprehends key content words/phrases in context
YES	NO	Retrieves simple factual information from short texts
YES	NO	Comprehends regular sound/symbol relationships
YES	NO	Sight reads key function words
YES	NO	Identifies genre of common texts
YES	NO	Identifies topic of simple text on familiar subject
YES	NO	Uses alphabetical indexes
YES	NO	Follows written instructions

(Source: Scarino, A. et al. 1988. Evaluation, Curriculum Renewal and Teacher Development. Australian Language Levels Guidelines Book 4. Canberra: Curriculum Development Centre.)

Table 16: Performance indicators

Content			
• Completion of activity	activity not completed		activity totally completed
Quality of performance			
Communication goals			
• comprehension of information (from interlocutor or text)	minimal comprehension		total comprehension
• intelligibility of response	minimally intelligible		totally intelligible
• quality of language resource:			
degree of accuracy (including grammar, vocabulary, pronunciation)	minimal accuracy		high accuracy
degree of fluency (speed and rate of utterance, ability to structure discourse)	minimal fluency		high fluency
range of expression (ability to go beyond stereotyped forms and to generate language)	limited range		good range
Sociocultural goals			
• sociocultural appropriateness	inappropriate		appropriate
• sociocultural knowledge	minimal knowledge		good knowledge
Learning-how-to-learn goals (including skills and strategies)			
• use of communication strategies	minimal use		effective use
• level of support required	strong reliance on support		no support required
General knowledge goals			
• knowledge of subject matter of the activity	minimal knowledge		good knowledge

(Source: Scarino, A. et al. 1988. Evaluation, Curriculum Renewal and Teacher Development. Australian Language Levels Guidelines Book 4. Canberra: Curriculum Development Centre.)

Table 18: General criteria for judging performance in activity-type 2

Activity-type 2	General criteria
Participate in social interaction related to solving a problem, making arrangements, making decisions with others, transacting to obtain goods, services, and public information (interacting and deciding)	**Conversation activities** • Did the learner succeed in solving the problem/making arrangements/ arriving at a decision/obtaining the particular goods or services? • Did the learner understand the information provided by others? • Were the learner's utterances intelligible? • Were the learner's utterances sufficiently accurate so as not to interfere with conveying meaning? • Were the learner's utterances appropriate to the sociocultural context? • Did the learner's responses cohere with the flow of the discussion? • Was the learner able to interact with others, take turns, maintain the conversation, generate questions, build on ideas? • Did the learner need help from others? • Did the learner provide information for the discussion? **Correspondence activities** • Did the learner complete the activity set? • Did the learner understand the information provided in the stimulus? • Was the learner's response intelligible? • Was the learner's response sufficiently accurate so as not to interfere with conveying meaning? • Was the learner's response appropriate to the sociocultural context? • Was the learner's response coherent? • Did the learner need support from the stimulus model or dictionary (if provided)?

Table 19: General criteria for judging performance in activity-type 3(a) & 3(b)

Activity-types 3a & 3b	General criteria
3a Obtain information by searching for specific details in a spoken or written text, and then process and use the information obtained (searching and doing) 3b Obtain information by listening to or reading a spoken or written text as a whole, and then process and use the information obtained (receiving and doing)	• Did the learner understand and extract the relevant information relating to the activity set? • Did the learner reproduce the information, as required by the activity? • Did the learner make an appropriate decision/choice/response on the basis of the information obtained • Was the learner's response intelligible? • Was the learner's response sufficiently accurate so as not to interfere with meaning? • Was the learner's response appropriate to the sociocultural context? • Was the learner's response coherent? • To what extent did the learner need support from others (interlocutor, or spoken or written text)?

Note: all macroskills are implied in these activity-types. Responses may be oral or written.