DOCUMENT RESUME

ED 365 150 FL 021 766

AUTHOR	Brindley, Geoff
TITLE	Defining Language Ability: The Criteria for Criteria.
PUB DATE	91
NOTE	27p.; In: Sarinee, Anivan, Ed. Current Developments in Language Testing. Anthology Series 25. Paper presented at the Regional Language Centre Seminar on Language Testing and Language Programme Evaluation (April 9-12, 1990); see FL 021 757.
PUB TYPE	Speeches/Conference Papers (150) -- Viewpoints (Opinion/Position Papers, Essays, etc.) (120)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Communicative Competence (Languages); Competency Based Education; *Criterion Referenced Tests; Evaluation Criteria; *Language Aptitude; *Language Proficiency; Language Research; Language Role; *Language Tests; Language Usage; Research Needs; Second Language Learning; *Second Languages; *Standardized Tests; Student Evaluation; Testing

ABSTRACT

Problems associated with criterion-referenced language testing are discussed in the context of both standardized proficiency testing and classroom assessment. First, different interpretations of criterion-referencing are examined. A range of approaches for defining criteria and performance levels in second language assessment are outlined, and some issues that have arisen in defining and applying these criteria are discussed, including the difficulties of defining the nature of proficiency and the failure of expert judges to agree on criteria. Finally, a discussion is given of research directions that might lead to language assessment criteria that incorporate multiple perspectives on learners' communicative needs and derive from empirical data on second language acquisition, variability in language use, and communicative competence. A 73-item bibliography is included. (MSE)

DEFINING LANGUAGE ABILITY: THE CRITERIA FOR CRITERIA

Geoff Brindley

INTRODUCTION

In recent years, there has been a move towards the wider use of criterion-referenced (CR) methods of assessing second language ability which allow learners' language performance to be described and judged in relation to defined behavioural criteria. This is in line with the concern among language testers to provide meaningful information about what testees are able to do with the language rather than merely providing test scores. However, while criterion-referencing has enabled language testers to be more explicit about what is being assessed, there are numerous problems associated with the development, interpretation and use of CR methods of assessment, to such an extent that the feasibility of true criterion-referencing has been questioned by some writers (eg. Skehan 1984, 1989).

This paper aims to illustrate and discuss the nature of these problems in the context of both standardized proficiency testing and classroom assessment. First, different interpretations of "criterion-referencing" will be examined. Following this, a range of approaches to defining criteria and performance levels in second language assessment will be outlined and some of the issues which have arisen in defining and applying these criteria will be discussed, including the difficulties of defining the nature of "proficiency" and the failure of expert judges to agree on criteria. Finally, research directions will be indicated that might lead to language assessment criteria which incorporate multiple perspectives on learners' communicative needs and which derive from empirical data on second language acquisition and use.

CRITERION-REFERENCING

The term "criterion-referenced" has been interpreted in a variety of ways in both general education and language learning. In their original formulation of the concept, Glaser and Klaus (1962: 422), in the context of proficiency measurement in military and industrial training, stated that

> *knowledge of an individual's score on a criterion-referenced measure provides explicit information as to what the individual can or cannot do*

Glaser (1963) described criterion-referenced assessment (CRA) thus:

> *The degree to which his achievement resembles desired performance at any specified level is assessed by criterion-referenced measures of achievement or proficiency. The standard against which a student's performance is compared when measured in this manner is the behaviour which defines each point along the achievement continuum. The term 'criterion', when used this way, does not necessarily refer to final end-of-course behaviour. Criterion levels can be established at any point in instruction as to the adequacy of an individual's performance. The point is that the specific behaviours implied at each level of proficiency can be identified and used to describe the specific tasks a student must be capable of performing before he achieves each of these knowledge levels. It is in this sense that measures of proficiency can be criterion-referenced.*

This early definition of CRA highlights several key elements which are reflected in various kinds of language assessment instruments: first, proficiency (here, interestingly, not distinguished very clearly from achievement) is conceived of as a continuum ranging from no proficiency at all to "perfect" proficiency; second, the *criterion* is defined as an external standard against which learner behaviour is compared; and third, levels of proficiency (or achievement) are linked to specific tasks.

CRITERION-REFERENCING IN LANGUAGE ASSESSMENT

In the context of language learning, CRA has number of different meanings (Skehan 1989: 5-6). In the first instance, it refers in a general sense to tests or assessments which are based on sampling of a behavioural domain and which make explicit the features of this domain. For example, in an oral interview, a testee might be given a score on a rating scale which contains the key aspects of performance (that is, the *criteria*) to be assessed such as fluency, appropriacy, accuracy, pronunciation, grammar etc. These criteria may then be described more fully in a band or level description. As Skehan (1984: 217) notes, such descriptions represent a set of generalised behaviours relating performance to external criteria (referred to by Jones 1985: 82 as the *performance criterion*), rather than a statement that would enable a yes/no decision to be made with respect to a testee's ability on a particular task.

As we have seen, CRA also carries a second meaning of a *standard* (criterion level) or cut-off point which may be defined with reference to some external requirement. In the context of language assessment, this might be exemplified by the "threshold level" set by the Council of Europe as a minimal level of functional language competence. Some writers, in fact, posit the existence of a constant and "natural reference point" for this external standard in the form of the native speaker (see, for example, Cziko 1983: 294).

Skehan (1989) also suggests a third sense in which CRA can be interpreted:

> *This is that the proficiency levels which are the basis for criterion-referencing are linked in some cumulative way to a course of development.*

This raises the issue of whether assessment criteria should take as their reference point what learners do, what linguists and teachers think learners do or what native speakers do. This point will be taken up later.

NORM-REFERENCING VERSUS CRITERION-REFERENCING

CRA is traditionally contrasted with *norm-referenced* methods of assessment which are meant to compare individual's performances relative to each other and to distribute them along the normal curve, not to establish the degree to which students have mastered a particular skill (Hudson and Lynch 1984: 172). Large-scale standardized examinations, in which students are given aggregate scores or grades for purposes of selection, certification or placement are probably the best-known example of norm-referenced assessment. An example of a norm-referenced approach from second language learning would be proficiency test batteries in which results are reported solely in terms of an overall score (a range of such tests is described by Alderson, Krahnke and Stansfield 1987).

According to some authors, however, the differences between norm-referenced assessment and CRA however, are not as great as conventionally imagined. Rowntree (1987: 185-6), for example, notes that criterion levels are frequently established by using population norms:

> *So much assessment that appears to be criterion-referenced is, in a sense, norm-referenced. The difference is that the student's performance is judged and labelled by comparison with the norms established by other students elsewhere rather than those established by his immediate fellow-students.*

There is an element of both norm- and criterion-referencing about the way in which proficiency descriptions are drawn up and interpreted. For example, one method of defining assessment criteria and performance descriptors for writing proficiency is to ask experienced teachers, without the aid of any explicit criteria, to rank learners in order of proficiency by sorting a set of writing scripts into piles representing clearly definable proficiency differences. Following this, the characteristic features of scripts at each level are discussed and these are then used to establish criteria and performance descriptors.

The level descriptions in proficiency scales, as numerous authors have pointed out (eg. Trim 1977, Skehan 1984), often contain norm-referenced terminology despite their claim to be criterion-referenced. Terminology such as "greater flexibility" or "fewer errors" relates the levels to each other instead of to the external standard which is supposed to characterise criterion-referencing. In terms of their actual use, as well, the descriptors may be interpreted in covertly norm-referenced ways. It is not unusual, for example, to hear teachers refer to a "good Level 1", a "slow Level 2" etc.

DEVELOPING CRITERIA AND DESCRIBING PERFORMANCE

Real world and classroom dimensions of CRA

CRA has both a real-world and a classroom dimension. In the development of a proficiency test aimed at assessing real-world language use, defining criteria involves operationalising the construct of proficiency -- in other words, specifying the skills and abilities which constitute the test developer's view of "what it means to know how to use a language" (Spolsky 1986). From the test specifications thus established, items are constructed and/or level/band descriptions written according to which performance will be rated. This is, of necessity, a time-consuming and rigorous process involving empirical studies of performance samples, consultation with expert judges and continuing revision of criteria and descriptors (see, for example, the descriptions by Alderson (1989) and Westaway (1988) of the way in which IELTS bands were derived).

In classroom CRA which is aimed at assessing learner achievement or diagnosing difficulties, the process of defining criteria and descriptors involves specifying the behavioural domain from which objectives are drawn, formulating a set of relevant objectives and establishing a set of standards by which learners' performance is judged. In many ways, this process replicates what is involved in operationalising the construct of proficiency, ie. that it involves specifying the nature of the domain to be assessed and breaking this down into its component parts. However, classroom CRA is likely to be less formal and may rely on implicit judgements on the teacher's part as to what constitutes the domain of ability which is assessed (Black and Dockrell 1984: 42-43).

It is worth noting at this point that the interpretation of "criterion" is slightly different, according to the purposes for which CRA is being carried out. Where learners' proficiency is being assessed in order to determine their capacity to undertake some real-world activity (eg. to exercise a profession), *criterion-referenced* is often taken to mean that their performance is compared against a "criterion level" of performance or a cut-score. They either reach the criterion or they don't. As Davies (1988: 33) notes, users of tests interpret all test results in a criterion-referenced way. A candidate's actual score is of less importance than the question: has the candidate attained the cut score or not?

In the classroom, however, the emphasis is slightly different. Here, the "criterion" against which learners' performance is assessed relates to a domain specification and a set of learning objectives. Attainment may be assessed in terms of mastery/non-mastery of these objectives (see, for example, Hudson and Lynch 1984; Hudson 1989). However, making a yes/no decision on whether mastery has been attained can be extremely difficult. In fact, the validity of the concept itself has been questioned (Glass 1978) and there are a multiplicity of competing views on appropriate standard-setting methods in CRA (see Berk 1986 for a comprehensive discussion of the relative merits of various methods). For this reason, classroom CRA is often more concerned with assessing learners' attainment on a scale of ability which represents varying degrees of mastery but is not necessarily linked to a "cut-score" (see Brindley 1989 for examples).

In terms of content, CR proficiency testing tends to focus on assessing tasks which replicate real life or from which inferences can be made to real-life performance. As far as classroom assessment is concerned, however, opinions differ on the question of whether CRA should be exclusively focussed on subsequent extra-classroom tasks or whether *any* valid objective can be assessed (Brown 1981: 7). If the latter view is accepted, then it would be possible to imagine situations in which CRA assessment did not concern itself with elements of learners' communicative performance (eg. if the syllabus were grammatically-based). CRA does not, in other words, necessarily mean communicative assessment. However, in the case of second language learners who have to use the language in society on a daily basis there are clearly arguments for accentuating methods of CRA which allow them to gain feedback on their ability to perform real-life tasks (see Brindley 1989: 91-120 for examples).

DEFINING CRITERIA

A variety of methods have been used by test developers and teachers to define assessment criteria and performance descriptors. These will be described below and some problems associated with each will be discussed.

Use existing criteria

The easiest way to define criteria and descriptors for language assessment is to use those already in existence. There is no shortage of models and examples. For proficiency testing, literally thousands of rating scales, band scales and performance descriptors are used throughout the world. An equivalent number of skills taxonomies, competency checklists, objectives grids etc, are available for classroom use.

Like tests, some proficiency scales seem to have acquired popular validation by virtue of their longevity and extracts from them regularly appear in other scales. The original scale used in conjunction with the Foreign Service Institute Oral Interview (FSI 1968), in particular, seems to have served as a source of inspiration for a wide range of other instruments with a similar purpose but not necessarily with a similar target group. Both the Australian Second Language Proficiency Rating Scale (ASLPR) (Ingram 1984) and the ACTFL Proficiency Guidelines (Hiple 1987) which aim to describe in the first case the proficiency of adult immigrants in Australia and in the second the proficiency of foreign language students and teachers in the USA, draw on the FSI descriptions.

Problems

Although proficiency scales have gained widespread acceptance over a considerable period of time and appear face-valid, it is very difficult to find any explicit information on how the descriptions were actually arrived at. Although some scales are claimed to be data-based (see, for example, Liskin-Gasparro (1984: 37) who states that the ACTFL guidelines were developed empirically), no information is made publicly available as to how the data were collected, analysed and turned into performance descriptors. This is despite the fact that in some cases claims are being made (if only by inference) to the effect that the descriptions constitute universal descriptions of second language development. Byrnes (1987), for example, claims that the ACTFL/ETS scale is built on a "hierarchy of task universals" .

Apart from their lack of empirical underpinning, the validity of rating scale descriptors (in particular the ACTFL/ETS Oral Proficiency Interview) has been contested on a number of other grounds. Some of the principal concerns which have been voiced can be roughly summarised as follows:

- the logic of the way levels are arrived at is essentially circular--"the criteria are the levels and vice versa" (Lantolf and Frawley 1985: 340). They cannot therefore be criterion-referenced in the accepted sense since there is no external standard against which the testee's behaviour may be compared.

- the incremental and lockstep nature of level descriptions fails to take into account the well documented variability and "backsliding" which occur in interlanguag (Pienemann, Johnston and Brindley 1988); nor can differential abilities in different "discourse domains" be accounted for (see Douglas and Selinker 1985, Zuengler 1989). In particular, the assumption that grammatical and phonological accuracy increases in a linear fashion is contradicted by evidence from second language acquisition studies which have shown systematic variability according to the learner's psycho-sociological orientation (Meisel et al. 1981); emotional investment in the topic (Eisenstein and Starbuck 1989); the discourse demands of the task (Brown and Yule 1989); desired degree of social convergence/divergence (Rampton 1987); planning time available (Ellis 1987); and ethnicity and status of interlocutor (Beebe 1983)

- not only are the performance descriptions covertly norm-referenced (see above), but also there is no principled relationship between co-occurring performance features which figure in the one level (Skehan 1984, Brindley 1986).

- it is very difficult to specify relative degrees of mastery of a particular skill with sufficient precision to distinguish clearly between levels. This is illustrated by Alderson's (1989: 11) comment on the development of the IELTS Speaking scales:

  For some criteria, for example pronunciation or grammatical accuracy, the difference in levels came down to a different choice of quantifiers and we were faced with issues like is 'some' more than 'a few' but fewer than 'several' or 'considerable' or 'many'. How many is 'many'?

- the essentially interactive nature of oral communication is inadequately represented due to the restriction of the possible range or roles which can be assumed by the non-native speaker (Lantolf and Frawley 1988; Raffaldini 1988; van Lier 1989).

- the descriptions are highly context dependent and thus do not permit generalisation about underlying ability (Bachman and Savignon 1986; Skehan 1989). Methods such as the oral interview confuse trait and method (Bachman 1988).

- in the absence of concrete upper and lower reference points, criterion-referencing is not possible. Bachman (1989: 17) points out that criterion-referencing requires the definition of the end points of an absolute scale of ability (so-called "zero" and "perfect" proficiency). Yet in practice, no-one has zero proficiency, since some language abilities are universal. Similarly, native speakers vary widely in ability, which makes the "perfect speaker" an equally tenuous concept.

Clearly the validity of the criteria on which proficiency descriptions are built is by no means universally accepted. However, the controversy surrounding the construct validity of proficiency rating scales and performance descriptors is merely a manifestation of the fundamental question that CRA has to face: how to define the domain of ability which is to be assessed, that is, language proficiency? Criterion-referencing depends on a very detailed and exact specification of the behavioural domain. But this amounts to asking the question posed by Spolsky (1986):

What does it mean to know how to use a language?

As far as proficiency testing is concerned, a definitive answer to this question is clearly not presently on the horizon, although detailed and testable models such as that proposed by Bachman (1990) offer some hope of describing more exactly the nature of communicative language ability. Meanwhile, in the context of classroom assessment, the move towards criterion-referencing continues. There is an increasing number of objectives-based assessment and profiling schemes derived from specification of real-life communicative needs which allow cumulative attainment to be monitored and documented in the form of profiles of achievement (see Brindley 1989: 91-111). These present a way of linking classroom assessment closely to real-world outcomes. However, objectives-based domain specifications also require the operationalization of the behaviour which forms the basis of the domain. As such, they are open to question on the same grounds as the proficiency descriptions described above. In addition, some testers would claim that performance testing associated with assessment of course objectives gives no information on underlying ability (Skehan 1989: 7).

The problem of domain specification is clearly far from being resolved. In the meantime, disagreement on the validity of criteria will no doubt continue, since there is as yet no description of language learning and language use on the basis of which universally agreed criteria could be drawn up.

Attacking the domain specification problem

Because of the limitations of context-dependent proficiency descriptions and the difficulties of relating these to an 'absolute' scale of ability, Bachman (1989) argues that the only way to develop adequate CR procedures for assessing communicative language proficiency is to attempt to clearly specify the abilities that make up language proficiency and to define scales or levels of proficiency which are independent of particular contexts, 'in terms of the relative presence or absence of the abilities that constitute the domain' rather than 'in terms of actual individuals or actual performance' (Bachman 1989: 256). An example of such a scale is given below.

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

Figure 1 Scales of ability in vocabulary and cohesion (Bachman and Palmer, 1983)

However such scales, too, are clearly fraught with problems as Bachman and Savignon (1986: 388) recognize when they admit the difficulty of 'specifying the degree of control and range in terms that are specific enough to distinguish levels clearly and for raters to interpret consistently'. The sample scales, in fact, manifest many of the same problems which arise in the design of more conventional proficiency rating scales. The terminology used is very imprecise and relativistic ('limited'; 'frequently'; 'confusing' etc) and in the absence of precise examples of learners' language use at each of the levels, problems of rater agreement would inevitably arise. In fact, since the levels do not specify particular contexts, structure, functions and so on, raters would not have any concrete criteria to guide them. The difficulties of reaching agreement between raters would, consequently, be likely to be even more acute.

Consult expert judges

Another commonly used way of producing criteria for proficiency testing is to ask expert judges to identify and sometimes to weight the key features of learner performance which are to be assessed. *Experienced teachers* tend to be the audience most frequently consulted in the development and refining of criteria and performance descriptions (eg. Westaway 1988; Alderson 1989; Griffin 1989). In some cases they may be asked to generate the descriptors themselves by describing key indicators of performance at different levels of proficiency. In others, test developers may solicit comments and suggestions from teachers for modification of existing descriptors on the basis of their knowledge and experience.

In ESP testing, *test users* may also surveyed in order to establish patterns of language usage and difficulty, including the relative importance of language tasks and skills. The survey results then serve as a basis for test specifications. This procedure has been followed in the development of tests of English for academic purposes by, *inter alia*, Powers (1986), Hughes (1988) and Weir (1983, 1988) and by McNamara (1989) in the construction of tests of speaking and writing for overseas-trained health professionals in Australia.

Problems

Who are the experts?

The idea of using "expert judgement" appeals to logic and common sense. However it poses the question of who the experts actually are. Conventionally it is teachers who provide "expert" judgements, although increasingly other non-teacher test users are being involved in test development. There are obvious reasons, of course, for appealing to teacher judgements. They are not difficult to obtain since teachers are on hand, they are familiar with learners' needs and problems, they are able to analyse language and they can usually be assumed to be aware of the purposes and principles of language testing, even though they may not always be sympathetic to it. Although less obviously "expert" in the sense of being further removed from the language learning situation and less familiar with linguistic terminology, test users who interact with the target group (such as staff in tertiary institutions or employers) can similarly be presumed likely to have some idea of the language demands which will be made on the testee and thus to be able to provide usable information for test developers.

But in addition to teachers and test users, it could also be argued that testees/learners themselves are "experts" on matters relating to their own language use and that their perceptions should also be considered in drawing up test criteria and specifications. Self-assessment based on learner-generated criteria is becoming increasingly common practice in classroom-based formative assessment and quite high correlations have been found between self-assessment and other external measures (Oskarsson 1989). However, learner perspectives have only recently begun to figure in proficiency test development (LeBlanc and Painchaud 1985; Bachman and Palmer 1988).

So-called "naive" native speakers constitute another "expert" audience whose perceptions could profitably be drawn on in establishing performance criteria. As Barnwell (1987) forcefully argues:

> *.....the domain of proficiency is outside the classroom not inside. We can (perhaps) leave achievement testing to the teachers and professional testers, but once we aspire to measure proficiency it becomes a question of vox populi, vox dei.*
> *Language is central to our humanity, and it is the most democratic and egalitarian attribute we share with our fellow man. Why then should we need 'experts' to tell us how well we speak? Thus it is not just an interesting novelty to contemplate the use of 'native' natives in proficiency testing and rating, it is a logical necessity which arises out of the nature of the thing we are trying to measure.*

Given that it is native speaker judgements of proficiency which may well determine the future of testees, it clearly important to investigate on what basis these judgements are made. As Clark and Lett (1988: 59) point out, comparing native speaker judgements with proficiency descriptors is one way of validating the descriptors in a non-circular way and of establishing the external criteria which have been lacking up to the present.

Data collection is resource-intensive

In order to establish valid performance criteria, an analysis of the testees' future domain of language use is clearly desirable. However, the collection of data for test construction purposes poses a number of logistical difficulties. From a practical point of view, the investigation of communicative needs is extremely resource-intensive, to such an extent that the practical constraints of data-gathering may end up jeopardizing the purpose for which the data are being gathered. (This same point had been made in relation to the rigorous needs assessment procedures which accompanied "target situation analysis" in ESP course development). An example is provided in a study by Stansfield and Powers (1989) aimed at validating the Test of Spoken English as a tool for the selection and certification of non-native health professionals and to establish minimum standards of proficiency. They state:

> *of necessity we asked for relatively global ratings, even for professionals and chose situations that would be representative and typical of those in which each professional might be involved. No attempt was made to specify all the many situations that might be encountered, nor was any effort made to designate highly specific tasks. We might have asked about the degree of speaking proficiency needed in the performance of surgical procedures, for example (in which oral proficiency might be critical) but time limitations precluded such detail. In addition in this study, we decided to consider neither other important dimensions of communicative competence (eg. interpersonal skills and other affective components) nor functions of language (eg. persuading or developing rapport with patients) that might be highly desirable in various medical situations.*

In only considering global proficiency, a course of action they were forced to take through lack of necessary resources, the researchers neglected the information which would be considered most essential by some (prospective patients is one group which springs to mind!) for test validity.

Precise information is difficult to elicit

An additional problem in consulting test users or "naive" native speakers in drawing up criteria for assessment is the difficulty of getting them to be sufficiently precise about situations of language use to provide usable information. Powers (1986), reporting on his attempts to elicit information from faculty members on university students' listening patterns, observes that:

> *the notion of analysing listening activities may have been "foreign" to many faculty members who were not involved intensely in language instruction or testing. In particular, such concepts as "discourse cues" and "non-verbal signals" may be somewhat far afield for faculty in non-language disciplines. Moreover, while the rating of such passive, non-observable skills as listening may be difficult generally, non-language oriented faculty may have even greater difficulty in determining when students encounter specific kinds of problems.*

Native speakers are not language analysts. Nor are most learners. It is hardly surprising, therefore, that the test users' perceptions of language needs tend to be stated in rather vague terms. This is exemplified by an examination by Brindley, Neeson and Woods (1989) of the language-related comments of 63 university supervisors' monitoring reports on the progress of foreign students. They found that the vast majority of the comments were of the general kind ("has problems with writing English"; "English expression not good"), though a few lecturers were able to identify particular goal-related skills ("has difficulty following lecturers-speak very fast").

In a similar vein, Weir (1988: 73), commenting on the development of a test specification framework for the TEEP test of English for academic purposes, notes that

> *There is a need for more precise methods for dealing with task dimensions than the pragmatic ones used in our research. We relied heavily on the judgements of teachers and other experts in the field, as well as on the results of small test administrations, to guide us on the appropriateness of task dimensions in the various constructs. Unless finer instruments are developed than these rather coarse subjective estimates, it is difficult to see how fully parallel versions of the test can ever be developed.*

Expert judgement may be unreliable

If expert opinion is to have any currency as a method of developing criteria, then one would expect that a given group of expert judges would concur, first on the criteria which make up the behavioural domain being assessed and second, on the allocation of particular performance features to particular levels. (Obtaining data in this way would be an integral part of construct validation). One would also expect that the group would be able to agree on the extent to which a test item was testing a particular skill and the level of difficulty represented by the item (agreement would constitute evidence for content validity).

Studies aimed at investigating how expert judgements are made, however, cast some doubt on the ability of expert judges to agree on any of these issues. Alderson (1988), for example, in an examination of item content in EFL reading tests, found that judges were unable to agree not only on what particular items were testing but also on the level of difficulty of items or skills and the assignment of these to a particular level. Devenney (1989) who investigated the evaluative judgements of ESL teachers and students of ESL compositions, found both within-group and between-group differences in the criteria which were used. He comments:

> *Implicit in the notion of interpretive communities are these assumptions: (1) a clear set of shared evaluative criteria exists, and (2) it will be used by members of the interpretive community to respond to text. Yet this did not prove to be the case for either ESL teachers or students*

Different people use different criteria

Non-teacher native speakers, teachers and learners themselves, by virtue of their different backgrounds, experiences and expectations, have different understandings of the nature of language learning and communication. As a result, they tend to use different criteria to judge language ability and thus to pay attention to different features of second language performance. Studies of error gravity, for example, have shown that native speakers tend to be less concerned with grammatical accuracy than teachers (particularly those who are not native speakers of the language taught (Davies 1983)). This highlights the difficulties of constructing assessment criteria and descriptors which can be consistently interpreted by different audiences.

It is interesting, and perhaps significant, to note in the context of this discussion that disciplines outside applied linguistics interpret "communication" or "communicative competence" quite differently and hence employ different criteria for assessment. Communication theorists, for example, accentuate criteria such as *empathy, behavioural flexibility and interaction management* (Wiemann and Backlund 1980) and emphasise the role of non-verbal aspects of communication. In other fields, such as organisational management, communicative ability is seen very much in terms of "getting the job done" and the success of communication is thus judged primarily in relation to how well the outcomes are achieved rather than on specific linguistic features (Brindley 1989: 122-23). McNamara (1987: 32) makes this point in relation to doctor-patient communication, noting that in the medical profession "there is a concern for the communication process in terms of its outcomes". He comments (1987: 47) that "sociolinguistic approaches to 'communicative ability' are indeed narrow, and narrowly concerned with language rather than communicative behaviour as a whole".

Two conclusions can be drawn from these observations. First, as McNamara (op. cit.) points out, we must be conscious of the limitations of the claims which can be made about the capacity of language tests to predict communicative ability (in the broader sense) in real-life settings. Second, if real-life judgements of communicative effectiveness are based on perceptions of people's ability to use language to complete a task satisfactorily, then it is worth trying to build this notion into assessment criteria. In this regard, the use of "task fulfilment" as a criterion in the IELTS writing assessment scales is a promising step in this direction (Westaway 1988).

Teachers will be teachers

Although teachers' judgements are frequently used as a basis for establishing assessment criteria, there is some evidence to suggest that the influence of their background and experience may be sufficiently strong to override the criteria that are given. For example, in a preliminary analysis of 12 videotaped moderation sessions of oral interviews conducted for the purposes of rating speaking ability at class placement in the Australian Adult Migrant Education Program, I have found a consistent tendency for teachers to:

- refer to criteria which are not contained in the performance descriptors at all, such as confidence, motivation, risk-taking capacity and learning potential.

- concentrate heavily on the assessment of some features of performance at the expense of others. In this case, more time was spent discussing the role of the grammatical accuracy than any other single factor, even though the descriptions being used did not provide detailed or specific comments on grammatical features.

- use diagnostically-oriented and judgemental "teacher language" in applying the criteria, such as:

  She seemed to be weak on tenses
  I was a bit concerned about her word order generally

> *her language was letting her down*
> *She's got weak tense forms, not sure of her prepositions and quite often leaves off a final-s*

Caulley et al (1988), report on a similar phenomenon in the context of the evaluation of common assessment tasks used in the Victorian senior secondary English examination:

> *in their discussions the teachers rarely or even referred to the specified criteria. Their assessments were largely global, the language abstract and rarely substantiated by reference to anything concrete:*

This was exemplified by comments such as

> *he's got communicative sense*
> *he's more sure of his material*
> *there's a lack of flow*
> *she hasn't crystallised her ideas*

They note that

> *teachers are involved with the growth and development of human beings through practice and in the end were shown to be neither willing nor able to divorce the performance of an action from those aspects of it such as intention, effort and risk, which make it one performed by a growing and developing human beings. They thus included in their assessment of students an estimate of the risk involved for the particular student to present as he or she did and something for the effort (or lack of effort) made in the preparation, although neither is mentioned in the guidelines.*

Although such non-linguistic factors do not conventionally figure as criteria in definitions of proficiency, it would appear that they are included by teachers, perhaps because they are perceived as part of their educator's role. Specific assessment criteria may be developed rigorously and clearly spelled out, yet the teachers appear to be operating with their own constructs and applying their own criteria in spite of (or in addition to) those which they are given. This tendency may be quite widespread and seems to be acknowledged by Clark and Grognet

(1985: 103) in the following comment on the external validity of the Basic English Skills Test for non-English-speaking refugees in the USA:

> *On the assumption that the proficiency-rating criterion is probably somewhat unreliable in its own right, as well as based to some extent on factors not directly associated with language proficiency per se (for example, student personality, ailigence in completing assignments etc) even higher validity coefficients might be shown using external criteria more directly and accurately reflecting language proficiency*

Further support for the contention that teachers operate with their own criteria is provided by a study carried out by Griffin (1989) who examined the consistency of the rating of IELTS writing scripts over time using a Rasch Rating scale model. An analysis of rater statistics revealed that

> *For assessment 1, most raters appeared to 'fix' the underlying variable. On occasion 2, however, few raters appeared to fix the variable. There appears to have been a change in the criteria or in the nature of the variable being used to assign scripts to levels. The original criteria used in the familiarisation workshop and reinforced in the training workshop do not seem to have been used for assessment 2. Unfortunately it was assumed that the criteria would remain the same and were in fact supplied to the raters.*
>
> (Griffin 1989: 10)

He comments that

> *raters seem to be influenced by their teaching background and the nature of the criteria used can differ from rater to rater. Consensus moderation procedures appear to have controlled this effect to some degree but not completely.*
>
> (Griffin 1989: 13)

CONCLUSION

From this review of CRA, it should be clear, as Skehan (1984: 216) remarks, that "criterion-referencing is an attractive ideal, but extremely difficult to achieve in practice". As we have seen, the criteria which are currently used may not reflect what is known about the nature of language learning and use and they may not be consistently interpreted and applied even by expert judges.

If the ideal of CRA is to be attained, it is necessary to develop criteria and descriptors which not only reflect current theories of language learning and language use but which also attempt to embody multiple perspectives on communicative ability. As far as the first of these requirements is concerned, Bachman and his colleagues have put forward a research agenda to develop operational definitions of constructs in Bachman model of communicative language proficiency and validate these through an extensive program of test development and research (see, for example, Bachman and Clark 1987; Bachman et al 1988; Bachman 1990). One of the main virtues of this model, as Skehan (1990) points out, is that it provides a framework within which language testing research can be organised. It is to be hoped that the model will enable language iesters to systematically investigate the components of language ability as manifested in tests and that the results of such research will be used to inform the specifications on which assessment instruments are based.

Second language acquisition (SLA) research can also make a contribution to the development of empirically-derived criteria for language assessment which reflect the inherent variability and intersubjectivity of language use. First, research into *task variability* of the type reported in Tarone (1989), Taroı e and Yule (1989) and Gass et al (1989a: 1989b) provides valuable insights into the role that variables such as interlocutor, topic, social status and discourse domain might exercise on proficiency. Investigation of factors affecting *task difficulty* might also provide a more principled basis for assigning tasks to levels, a major problem in CRA. A number of testable hypotheses are outlined by Nunan (1989).

Second, SLA research could also provide much-needed information on the factors which influence native speaker perceptions of non-native speakers' proficiency. There is already a considerable literature on the overall communicative effect of non-native speaker communication (eg Albrechtsen et al 1980; Ludwig 1982; Eisenstein 1983) and error gravity (eg James 1977; Chastain 1980; Davies 1983). However such studies have tended to examine the effects of particular discourse, phonological, syntactic or lexical features on comprehensibility and/or irritation, rather than relating them to perceptions of proficiency. Studies conducted with a specific focus on proficiency would assist in the creation of performance criteria which reflect those used in real life. Information of this kind is of critical importance since in many cases, it is the judgements of native speakers that will determine the future of language learners, not so much those of teachers. At the same time, it is important to try to establish to what extent non-linguistic factors such as personality, social status, ethnicity, gender etc affect judgements of proficiency and the extent to which these factors can be related to linguistic ones (Clark and Lett 1987).

Third, research into the nature of *developmental sequences* in learner language gives an indication of the grammatical elements of language which can realistically be expected for production at different stages and thus provides a basis for establishing assessment criteria which are consistent with the regularities of language development (Pienemann et al 1988). In addition, since the multi-dimensional model of second language acquisition described by Pienemann and Johnston (1987) makes strong predictions concerning the processing demands made by different linguistic elements on learners, it should be possible to incorporate these predictions into concrete hypotheses concerning task difficulty which can be empirically investigated.

Thus far I have sketched out the kinds of research that might contribute to the development of better criteria. As far as the *interpretation* of the criteria is concerned, however, it would be naive to imagine that different judges will not continue to interpret criteria idiosyncratically. As Messick (1989) says:

> *....expert judgement is fallible and may imperfectly apprehend domain structure or inadequately represent test structure or both.*

Agreement between testers can be improved by familiarisation and training sessions in which raters, as Griffin (1989) reports. But there is always the possibility that agreement might conceal fundamental differences. As Barnwell (1985) comments:

> *raters who agree on the level at which a candidate can be placed may offer very different reasons for their decisions*

Given, as we have seen, that different judges may operate with their own personalized constructs irrespective of the criteria they are given, it would be a mistake to assume that high inter-rater reliability constitutes evidence of the construct validity of the scales or performance descriptors that are used. In order to provide such evidence, empirically-based investigation of the behavioural domain itself has to be carried out, as I have indicated above. At the same time, studies requiring teachers, learners and native speakers are to externalize the criteria they (perhaps unconsciously) use to judge language ability would help to throw some light on how judgements are actually made by a variety of different audiences and lead to a better understanding of the constructs that inform the criteria they use. The procedures used in the development of the IELTS band scales as reported by Westaway (1988), Alderson (1989), Griffin (1989) offer the possibility of building up a useful data base in this area.

Finally, in the context of classroom CRA, the time is ripe to explore the feasibility of incorporating communicatively-oriented CRA into the teaching and learning process. In the field of general education, the results of research into the development of CR instruments for classroom use indicates that the problems of domain specification described in this paper may not be as intractable as they are sometimes portrayed (Black and Dockrell 1984). Numerous CR schemes for formative assessment and profiling are in existence in general education the United Kingdom and Australia (see Brindley 1989 for an overview) and appear to be quite adaptable to second language learning situations. The use of CR methods of assessing achievement based on communicative criteria would not only help to link teaching more closely to assessment, but also would allow for closer involvement of learners in monitoring and assessing their progress.

ACKNOWLEDGEMENT

I would like to thank Charles Alderson for his helpful comments on an earlier version of this paper.

REFERENCES

Alderson, J C. 1989. Bunds and scores. Paper presented at IATEFL Language Testing Symposium, Bournemouth, 17-19 November.

Alderson, J C. 1988. Testing reading comprehension skills. Paper presented at the Sixth Colloquium on Research in Reading in a Second Language. TESOL, Chicago, March 1988.

Alderson, J C., K Krahnke and C W Stansfield (Eds.) 1987. Reviews of English Language Proficiency Tests. Washington: TESOL.

Bachman, L F. 1988. Problems in examining the validity of the ACTFL oral proficiency interview. Studies in Second Language Acquisition, 10, 2, pp. 149-164.

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

Bachman, L F. 1990. Fundamental Considerations in Language Testing. Oxford: Oxford University Press.

Bachman, L F., and S Savignon 1986. The evaluation of communicative language proficiency: a critique of the ACTFL oral interview. Modern Language Journal, 70, 4, pp. 380-390.

Bachman, L F., and J L D. Clark 1987. The measurement of foreign/second language proficiency. Annals of the American Academy of Political and Social Sciences, 490, pp. 20-33.

Bachman, L F., and A S Palmer 1983. Oral interview test of communicative proficiency in English. MS.

Bachman, L F., and A S Palmer 1988. The construct validation of self-ratings of communicative language ability. Paper presented at Tenth Annual Language Testing Research Colloquium, University of Illinois at Urbana-Champaign, March 5-7.

Beebe, L. 1983. Risk-taking and the language learner. (In) Classroom-Oriented Research in Second Language Acquisition, H. Seliger and M. Long (Eds.) Rowley, Massachusetts: Newbury House.

Berk, R A. 1986. A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56, 1, pp. 137-172.

Black, H D and W B Dockrell 1984. Criterion-Referenced Assessment in the Classroom. Edinburgh: Scottish Council for Research in Education.

Brindley, G. 1986. The Assessment of Second Language Proficiency: Issues and Approaches. Adelaide: National Curriculum Resource Centre.

Brindley, G. 1989. Assessing Achievement in the Learner-Centred Curriculum. Sydney: National Centre for English Language Teaching and Research.

Brindley, G S Neeson and S Woods 1989. Evaluation of Indonesia-Australia Language Foundation Assessment Procedures. Unpublished manuscript.

Brown, S. 1981. What do they know? A Review of Criterion-Referenced Assessment. Edinburgh: Her Majesty's Stationery Office.

Byrnes, H. 1987. *Proficiency as a framework for research in second language acquisition. Modern Language Journal, 71, 1, pp. 44-49.*

Caulley, D., Orton, J., and L Claydon 1988. Evaluation of English oral CAT. Melbourne: Latrobe University.

Chastain, K. 1980. Native speaker reaction to instructor-identified student second language errors. Modern Language Journal, 64, pp. 210-215.

Clark, J L D and J Lett 1987. A research agenda. (In) Second Language Proficiency Assessment: Current Issues, P Lowe and C W Stansfield (Eds.), Englewood Cliffs: Prentice-Hall, pp. 53-82.

Clark, J L D and A Grognet 1985. Development and validation of a performance-based test of ESL 'survival skills". (In) Second Language Performance Testing, P Hauptman, R LeBlanc and M Wesche (Eds.). Ottawa: Ottawa University Press.

Cziko, G. 1983. Psychometric and edumetric approaches to language testing. (In) Issues in Language Testing Research, J W Oller (Ed.), Rowley, Massachusetts: Newbury House, pp. 289-307.

Davies, E. 1983. Error evaluation: the importance of viewpoint. English Language Teaching Journal, 37, 4, 304-311.

Devenney, R. 1989. How ESL teachers and peers evaluate and respond to student writing. RELC Journal, 20, 1, pp. 77-90

Dickinson, L. 1987. Self-Instruction in Language Learning. Cambridge: Cambridge University press.

Douglas, D and L Selinker 1985. Principles for language tests within the 'discourse domains' theory of interlanguage: research, test construction and interpretation. Language Testing, 2, 2, pp. 205-226.

Eisenstein, M. 1983. Native-speaker reactions to non-native speech: a review of empirical research. Studies in Second Language Acquisition, 5, 2, pp. 160-176.

Eisenstein, M and R Starbuck. 1989. The effect of emotional investment on L2 production. (In) Gass et al (Eds.). 1989b.

Ellis, R. 1987. Interlanguage variability in narrative discourse: style-shifting in the use of the past tense. Studies in Second Language Acquisition, 9, 1, pp. 1-20.

Foreign Service Institute 1968. Absolute Language Proficiency Ratings. Washington, D C; Foreign Service Institute.

Gass, S., C Madden, D Preston and L Selinker (Eds) 1989a. Variation in Second Language Acquisition: Discourse and Pragmatics. Clevedon, Avon: Multilingual Matters.

Gass, S., C Madden, D Preston and L Selinker (Eds) 1989b. Variation in Second Language acquisition: Psycholinguistic Issues. Clevedon, Avon: Multilingual Matters.

Glaser, R., and D J Klaus 1962. Assessing human performance. (In) R Gagne (Ed.), Psychological Principles in Systems Development., New York: Holt, Rinehart and Winston.

Glaser, R. 1963. Instructional technology and the measurement of learning outcomes. American Psychologist, 18, pp. 519-521.

Glass, G V. 1978. Standards and criteria. Journal of Educational Measurement, 15, pp. 237-261.

Griffin, P E. 1989. Latent trait estimates of rater relibility in IELTS. Paper presented at Fourteenth Annual Congress of the Applied Linguistics Association of Australia, Melbourne, September 1989.

Hiple, D. 1987. A Progress report on the ACTFL proficiency guidelines 1982-1986. (In) Defining and Developing Proficiency, H Byrnes and M Canale (Eds.), Lincolnwood, Illinois: National Textbook Company.

Hudson, T. 1989. Mastery decisions in program evaluation. (In) The Second Language Curriculum Curriculum, R K Johnson (ed.), Cambridge: Cambridge University Press.

Hudson, T and B Lynch 1984. A criterion-referenced approach to ESL achievement testing. Language Testing, 1, 2, pp. 171-201.

James, C V. 1977. Judgements of error gravity. English Language Teaching Journal, 31, 2, pp. 175-182.

Jones, R L. 1985. Some basic considerations in testing oral proficiency. (In) New Directions in Language Testing, Y P Lee, A C Y Y Fok, R Lord and G Low (eds.), Oxford: Pergamon, pp. 77-84.

Lantolf, J P and W Frawley 1985. Oral proficiency testing: a critical analysis. Modern Language Journal, 69, 4.

Lantolf, J P, and W Frawley 1988. Proficiency: understanding the construct. Studies in Second Language Acquisition, 10, 2, pp. 181-195.

LeBlanc, R, and G Painchaud 1985. Self-assessment as a second language placement instrument. TESOL Quarterly, 19, 4, pp. 11-42.

Liskin-Gasparro, J. 1984. The ACTFL guidelines: a historical perspective. (In) Teaching for proficiency: The organizing principle, Lincolnwood, Illinois: National Textbook Company.

McNamara, T F. 1987. Assessing the Language Proficiency of Health Professionals. Recommendations for Reform of the Occupational English Test. Parkville, Victoria, University of Melbourne: Department of Russian and Language Studies.

McNamara, T F. 1989. ESP testing: general and particular. (In) Language, Learning and Community, C N Candlin and T F McNamara (eds.), Sydney: National Centre for English Language Teaching and Research, pp. 125-142.

Meisel, J, H Clahsen and M Pienemann 1981. On determining development stages in second language acquisition. Studies in Second Language Acquisition 3, pp. 109-135.

Messick, S. 1989. Meaning and values in test validation: the science and ethnics of assessment. Educational Researcher, 18, 2, pp. 5-11.

Nunan, D. 1989. Designing Tasks for the Communicative Classroom. Cambridge: Cambridge University Press.

Oskarsson, M. 1984. Self-Assessment of Foreign Language Skills. Strasbourg: Council of Europe.

Oskarsson, M. 1989. Self-assessment of language proficiency: rationale and applications. Language Testing, 6, 1, pp. 247-259

Pienemann, M, and M Johnston 1987. Factors influencing the development of language proficiency. (In) applying Second Language Acquisition Research, D Nunan (ed.), Adelaide: National Curriculum Resource Centre, pp. 45-141.

Pienemann, M., M Johnston and G Brindley 1988. Constructing an acquisition-based assessment procedure. Studies in Second Language Acquisition, 10, 2, pp. 217-243.

Powers, D E. 1986. Academic demands related to listening skills. Language Testing, 3, 1, pp. 1-38.

Powers, D E and C W Stansfield 1989. An approach to the measurement of communicative ability in three health professions. (In) Working with Language, H Coleman (ed.), Berlin: Mouton de Gruyter, pp. 341-366.

Raffaldini, T. 1988. The use of situation tests as measures of communicative ability. Studies in Second Language Acquisition, 10, 2, pp. 197-216.

Rampton, B. 1987. Stylistic variability and not speaking 'normal' English: some post-Labovian approaches and their implications for the study of interlanguage. (In) Second Language Acquisition in Context. R Ellis (Ed.), Englewood Cliffs: Prentice Hall, pp. 47-58.

Rowntree, D. 1977. Assessing Students: How Shall We Know Them? London: Harper and Row.

Skehan, P. 1984. Issues in the testing of English for specific purposes. Language Testing, 1, 2, pp. 202-220.

Skehan, P. 1988. State of the art: language testing. Part 1. Language Teaching. 21, 2, pp. 211-221.

Skehan, P. 1989. State of the art: language testing. Part 2. Language Teaching, 22, 1, pp. 1-13.

Skehan, P. 1990. Progress in language testing: the 1990s. Revised version of plenary address to IATEFL Language Testing Symposium, Bournemouth, 17-19 November 1989.

Spolsky, B. 1985. What does it mean to know how to use a language? An essay on the theoretical basis of language testing. Language Testing, 2, 2, pp. 180-191.

Tarone, E and G Yule 1989. *Focus on the Language Learner.* Oxford: Oxford University Press.

Tarone, E. 1988. *Variation in Interlanguage.* London: Edward Arnold.

Trim, J L M. 1977. *Some Possible Lines of Development for an Overall Structure for a European Unit/Credit Scheme for Foreign Language Leaming by Adults.* Strasbourg: Council of Europe.

van Lier, L. 1988. Reeling, writhing, fainting and stretching in coils: oral proficiency interviews as conversation. *TESOL Quarterly,* pp. 489-508.

Westaway, G. 1988. Developments in the English Language Testing Service (ELTS) M2 writing test. *Australian Review of Applied Linguistics,* 11, 2, pp. 13-29.

Wiemann, J M., and P Backlund 1980. Current theory and research in communicative competence. *Review of Educational Research,* 50, 1, pp. 185-199.

Wier, C J. 1988. The specification, realization and validation of an English language proficiency test. (In) *Testing English for University Study,* A Hughes (Ed.), London: Modern English Publications and the British Council.

Zuengler, J. 1989. Performance variation in NS-NNS interactions: ethnolinguistics difference or discourse domain? (In) Gass et al (Eds) 1989a pp. 228-243.