Is Proficiency Always Achievement?[1]

Alan Davies

I begin by trying to give a sense of the excitement of working in language testing and at the same time suggest what is to me its value.

A good place to start is with this well-known religious test from the Book of Judges (12/4) :

> *'So he (Jeptha) mustered all his Gileadites and fought against the men of Ephraim and the Gileadites defeated them. They had taunted the men of Gilead with being fugitives from Ephraim, living in territory that belonged to Ephraim and Manasses. So now these men of Gilead seized the fords of Jordan by which the Ephraimites must needs pass on their way home; and when any of Ephraim's men came up asking for passage, they would ask him, Art thou from Ephraim? Not I, would be his answer. Then they would bid him say the word Shibboleth, which means an ear of corn; and he would answer Sibboleth, pronouncing the word amiss. So then, without more ado, they would take him down to Jordan ford and slay him; there were then 42,000 men of Ephraim who then perished.'*

What a marvellous test that seems to be! Just one question, setting one decisive cut-off, gives enough information to separate the sheep from the goats. But in fact as a language test it is quite useless, too powerful, making a distinction only between groups, those who know perfectly and those who do not, and not taking any account of the individual and his/her qualities.

At the opposite extreme is the danger of certain special-purpose tests. Some years ago, I was involved in work of this kind with the English Language Testing Service (ELTS test, an earlier version of the current IELTS test. In that test a serious attempt was made to provide for specialist language use by the provision of tests in the language of subject areas such as medicine. The problem that arose

---

1 A version of this paper was given at the Fourth ELICOS Association Educational Conference at Monash University, Clayton, Victoria, Australia in August 1991.

was that the doctors taking the test, professionals who were seeking admission to postgraduate study in one of the medical colleges in the U.K., would often say on seeing the medical test. 'But this isn't my medicine. I'm a surgeon not a physician' or 'I'm a pediatrician not a gynaecologist'. The chimaera-like problem we found in this attempt to provide for specialist knowledge is that the logic leads to a different test for every candidate; only then are you being 'fair' to everybody.

That again is a hopeless quest. Neither a shibboleth-like group test nor a special purpose individual test serves any serious purpose as a language test. What is needed is something in the middle which is suitable for a group and at the same time does take account of individual differences.

With that in mind what most language testing is really about is the attempt to balance these demands, the group and the individual. That means taking some account of context (so that it's not all shibboleth) and aiming at a common scale (so it's not over-specialised purpose).

The skills called on therefore in language testing, in addition to the calibration skills which are necessary in order to provide for reliability (or consistency), are those of sampling, which is, I suggest, very similar to what the language teacher and indeed the linguist is called on to exercise. In all cases, tester, teacher and linguist, an attempt is being made to provide in a very short space — a test, a text-book or set of lessons, a grammar book — a summary of the whole of the language or language area under target. The language tokens that are presented in all these cases are only useful if in some way they are central to the language. The greatest burden falls more on the test than on the text-book or the linguist's grammar because the time available for test administration is very short. Public demands for a proficiency score often expect a very short test of, say, an hour at most to be adequate. But that is close to shibboleth country.

The sampling must therefore focus on the indicators of more general skill. In our Language Testing Centre (LTC) in the University of Melbourne one of our projects involves the construction of an instrument for measuring the English proficiency of non Australian trained teachers of science and maths. In this project we find

ourselves necessarily involved in negotiating among the many variables that influence the intelligibility of a teacher in the classroom, asking ourselves first of all to what extent the problem, if indeed there is a problem, is a language one, rather than a cultural or an attitudinal or a teaching style one, and then if it is a language problem, what are the salient features that trigger in the minds of the teachers' colleagues, their supervisors and trainers as well as in the pupils' and their pupils' parents', a sense of positive or negative comprehension.

I have mentioned culture. The relation of language and culture is an old song. But it becomes a practical issue in testing, as we are finding in our work on semi-direct tests of Japanese proficiency for the tourist and hospitality industry. We have encountered the same issue — clearly it won't go away — in the evaluation we did for the New South Wales Basic Skills Tests of Numeracy, and in our work on the Australian Language Certificates. The issue in all these cases is what is meant by proficiency — at of course different levels of learning. That is what does it mean to know a language, a knowledge which, as we see it, encompasses cultural awareness.

The task we set ourselves is to come up with as valid a language sample as we can and for that what is needed is of course expertise in the language. Here then language testers need the advice and the support of expert informants in the language under test, who can also help disentangle in the culture what is essential and generalisable from what is contingent. A test which is too culture loaded takes us back to the specialist purpose individual-focussed test, and I have argued that that does not tell us anything about an individual in terms of a common scale.

The responsibility on test constructors is a heavy one, not usually (fortunately) as heavy as on the men of Gilead. It is therefore of consequence that decisions about sampling, which involves the separating of those aspects of culture which are essential from those which are contingent, are taken with care and with a proper seriousness about the information that the test seeks to provide. With its involvement in certification and gate-keeping, testing is concerned with job opportunities and indeed with life chances.

For that reason, that in the sampling it is important to get at the underlying skills, it may not always be the case that the most direct

test is the best (or most valid) way of doing so. After all, the shibboleth test was very direct, full of what is sometimes called face validity, and yet how many of the 42,000 men of Ephraim were in reality migrants who hadn't learnt to say sibboleth? Sometimes a more devious approach may be preferable, for as Polonius said to Reynaldo:

> *And thus do we of wisdom and of reach,*
> *With windlasses and with assays of bias,*
> *By indirections find directions out.*
>   (Shakespeare, Hamlet 2/1, Polonius to Reynaldo)

· The title of my paper 'Is Proficiency Always Achievement?' deliberately brackets achievement and proficiency. Any attempt to distinguish these concepts must come to grips with the basic issue in all discussions of language ability, which is to what extent the concept of a single or unitary or general language ability is tenable. We see this issue emerge in various guises, in language testing the argument about general language proficiency or the unifactorial versus the multifactorial positions; in language pedagogy the general versus specific purpose language teaching (and presumably learning); and in second language acquisition in universal grammar as against the particular grammars of individual languages.

The tradition of distinguishing clearly between achievement (or attainment) and proficiency is a convenient one. Proficiency, it is suggested, is general, achievement specific and local; proficiency is theoretical or theory based; achievement is syllabus or materials or curriculum based, parasitic, in the sense that achievement information describes the learning of a single programme; while proficiency is free standing and describes learning in some absolute sense. From this point of view achievement is dependent through the syllabus and materials on some proficiency construct.

However, this clear-cut definition has been questioned. As Brindley (1989) and Bachman (1990), among others, have pointed out, an achievement test is often used as if it were a proficiency test, or rather it is used as a general indication of learning; equally, a proficiency test is difficult to distentangle fully from the circumstances of its use. In the first place, apparently similar performance on as robust a test as TOEFL (or to a lesser extent IELTS) can be shown to vary in terms of factors such as mother tongue. In the

second place, on a proficiency scale such as ASLPR the criteria influencing the levels allocated to different groups (eg a group of work place adults, groups of high school students or university postgraduates or ELICOS students) will not be identical. In other words, what count as criteria for a level in one context will not be the same as in another.

What is omitted in such arguments is the requirement of validity. In its weak version validity emphasises the importance of the claims a test makes. Here the claims of a so-called proficiency test such as IELTS or TOEFL are in fact likely to be more modest than the claims of a proficiency scale such as the Australian Second Language Proficiency Scales (ASLPR). That is to say that the IELTS (for example) insists that it is intended for academic purposes; the ASLPR and similar scales on the other hand seem to claim to be universally applicable, and by so doing overreach themselves: 'because the ASLPR was designed to measure general language proficiency it can be used for a whole variety of purposes for which a statement is required about a learner's proficiency in General English or in any of the four macroskills' (Ingram 1982:14). Even IELTS seems uneasy about spanning all forms of academic English; difficult enough to cope with various academic disciplines; much harder to take account of differing academic levels. The ASLPR seems to claim wide serviceability.

If we admit the strong version of validity we accept that a test cannot be seriously considered without a certification explicitly stating its claim to be valid. That is where validity can remove a test out of its context and relate it to a more general concern for ability. To some extent this is a sampling claim by which the good (that is valid) test, which contains only some part of what proficiency is thought to be, reflects all proficiency. Note that this claim is also true for the achievement type of test, in the sense that the most classroom bound test can never test all that is to be tested. That is to say that an achievement test is still extrapolating beyond the test — to what? to everything that has been learnt, yes, but also to what everything that has been learnt also stands for, to a whole language area or level, eventually to the more general case of proficiency. Equally, if this argument holds for a test such as IELTS or TOEFL, it must also hold for a proficiency scale such as the ASLPR. Here the rater extrapolates from the sample of behaviour

(for s/he can observe only a sample in any task based test situation) to the descriptors on the scale bands.

We might therefore suggest that the categorial difference between a test and a scale is that the test measures language behaviour without telling us what it means; the scale tells us what it means without helping us measure it. Is it possible to combine the two? No doubt there is always an implicit assumption that for a scale there is an appropriate test and vice versa. If the attempt is made, as appears to be the case with IELTS, to be more explicit, that, as Alderson points out when discussing the different audiences for scales, creates its own problems. Commenting on the IELTS Band Scales for Reading, Alderson writes 'the production and public availability of band descriptors commits the test developer to a clearly untenable assertion' ie that future parallel tests 'measure "the same thing" in a highly specific way' (1991:76).

Validity also allows us to concentrate on test purposes and uses rather than on test definitions. A basic testing question is what is the purpose of a test, what is it used for. As we shall see that is also to some extent a question about a test's validity.

Let us examine the uses of achievement and proficiency tests alongside two other test uses, that of the aptitude test and the diagnostic test (Davies 1977). The distinctions we make are in terms of time and subject-matter. Achievement tests are concerned with assessing what has been learned of a known syllabus within a school or a total educational system. These include the typical external school examinations (VCE), the university degree exams, even, by extension, the RSA teachers' exams and the Cambridge so-called proficiency exams. What is at issue is control of certification and the existence of a known and stated syllabus but as will be obvious the overlap with proficiency is considerable. At the extremes we are distinguishing between, say, an achievement test of the memorisation of a 10 item vocabulary list and a proficiency test of the whole of the language. In between there is more and less control over what is made explicit for the test: the more explicit the specification for the learning, the more achievement-like, the more the test is being used as an achievement test. The less explicit the more proficiency-like the test.

It is common to describe proficiency tests in terms of the assessment of what has been learned of an unknown syllabus, that is to say that the proficiency test is open to all comers. Courses are set up and text-books written, even for the most secure tests. That is why TOEFL teaching is as big an industry as TOEFL testing. Note that the claim I make here is precisely about test uses, that when a prestigious proficiency test (such as the Cambridge Proficiency Exam) is made the goal of teaching courses etc what changes is the role of the test, its use, as we are suggesting. That is why, in my view, the achievement-proficiency distinction is best seen in terms of use.

There is another characteristic of proficiency tests, this time distinguishing them from both achievement and aptitude tests, that is in their relation to future needs. While an achievement test is concerned solely with the information it provides about past instruction, and therefore with backwash — deliberately so — a proficiency test rejects control over the past but is hugely concerned with future learning. So of course is an aptitude test but while the aptitude test is concerned with predicting language learning, a proficiency test is concerned with predicting adequate control over language skills for an extralinguistic purpose (enough English to study engineering, enough Japanese to work as a tour guide and so on). The purpose of an aptitude test is therefore to determine whether a learner has sufficient aptitude or language learning ability to succeed in learning one or other language. In this sense therefore we may speak of proficiency in language X for job Y; and aptitude for language X for itself.

Diagnostic tests differ from the others in that they relate to the use of the information obtained and to the absence of a skill in the learner. Achievement, proficiency and aptitude are all concerned with both use and skill. A diagnostic test is a use made by a teacher of the information provided from the presence or absence of one or part of one of the skills (can't form the negative, ask questions, listen with comprehension to real-time speech etc). A diagnostic test may be constructed for itself or it may be an additional use made of an achievement or proficiency test. If it is specially constructed it could be argued that some element of learner's skill or rather absence of skill is involved because the tester is concerned with discovering what might be termed non-achievement. Like the proficiency test the diagnostic test may look before and after; before

to find out what is wrong with the previous learning, after in order to do something about it in the form of remedial work.

Diagnostic tests, in spite of their obvious utility, have been slow to appear, perhaps because of the time and length needed for adequate diagnosis. There is some optimism now that, with the advent of computer based language testing, computer adaptive tests may make good this need since they can provide extremely rapid and personalised access to a very large item bank (see for example Corbel 1990).

In terms of validity, achievement tests are solidly grounded in content validity; they justify their make-up on the basis of the materials and syllabus on which they are based. Proficiency and aptitude tests both make use of predictive and construct validity; they both measure their effectiveness in terms of success in their indications of those who will succeed and fail; and they are both related to a model or theory which provides an explanation of why the test is constructed in the way it is.

I turn now to a consideration of the relation between proficiency scales and tests. The increasing use of proficiency scales in language assessment has both positive and negative aspects. On the positive side they are authentic examples of language in use; there is no gap between what Bachman calls 'the criterion of proficiency and the definition of authenticity' (1990:409). Because such procedures are typically direct authenticity comes as it were free and does not have to be appealed to or claimed elsewhere. It is therefore often argued on behalf of such techniques that they have built-in validity. On the negative side it must surely be pointed out that all tests (not just indirect or semi-direct ones) lack authenticity. They are all simulations of real life rather than real life itself. What this of course suggests is that it is the job of assessment not so much to replicate real life (because by definition that cannot be done or when done is always partial and potentially biased) but to reflect language learning abilities and to sample real life situations rather than to collect them. The old example from the testing of reading makes the point forcibly: it is surely clear that when we test someone's literacy on a text we have no serious interest in that particular text. What we are interested in is the learner's ability to read texts. How important therefore it is that the text used for the test, and the tasks required in the test should be adequate samples

of texts at the appropriate reading level and of the tasks required in that reading.

Assessment which makes use of proficiency scales typically uses the interview test as a means of sampling language data which can then be related to the scale. Interviews are said — as I have hinted above — to provide direct entry into the speaker's language ability. But they are also notoriously unreliable. There are, of course, ways of training judges and ways of pooling judgements but the extent to which they can be made less unreliable is a function of the amount of training and time and money that are available. In other words *plus* reliable means *minus* practical. Which is to say that the nearer the test approximates to real life (the longer it is, the more assessors there are and so on) the more validity and the more reliability it will have. But as it approximates to real life it ceases to be a test and becomes precisely what it is supposed to be predicting, which is real life use.

Direct tests make more claims on reality than they do on testing. They confuse the criterion with the test. A compromise is needed between the claim of directness and the requirements of testing. Bachman (1990) describes such a compromise in his useful discussion of face validity and of direct tests (in his view direct tests are basically attempts to embody face validity, a concept he dismisses as not serious either academically or pedagogically). If face validity (and therefore direct tests) have any respectability it can only be in single settings. Direct tests have no generalisability beyond those single settings. Since the purpose of a test is to provide a sample of authentic language behaviours it must provide information about the abilities that underlie language performance in real life situations not just in that performance itself. A frontal assault on truth may be neither practical nor politic. I remind you of the earlier Hamlet quotation : 'by indirections find directions out'.

There is in direct tests, even the interview, the ever-present danger of routinisation (as indeed there is in the analogue, communicative language teaching). A good example of the danger can be found in the so-called *neck verse*. In the Middle Ages in England (and perhaps elsewhere) the Church and the state had separate legal systems so that the clergy were not subject to the jurisdiction of the state. Since the Church had less severe punishment (notably there was no capital punishment) at a time when very many offences were

punished by death by the state, it was obviously to a person's advantage to be tried for an offence by the Church by claiming 'benefit of clergy'. This required summary proof of being a cleric. Since that was not an easy matter, appellants were required to prove their clerisy to the state officials by reading a verse from the Bible. That was a direct test and no mistake! However, since they often had to read (in Latin of course) to people who were themselves either illiterate or only partially literate the practice grew up of always being required to read the same verse (from Psalm 51), the so-called 'neck verse', the one that begins : *'Miserere mei, deus'* ('Have mercy upon me, O God, after thy great goodness : according to the multitude of thy mercies, do away my offences'). Such routinisation led, as is obvious, to the abuse by many who were not clerics claiming the right to be tried by the Church for an offence.

But that is the danger inherent in all direct tests, that in order to be fair, to avoid subjectivity, the test becomes more and more routine as time goes on and eventually as little like real life as the most indirect test but without its special claim to be a sample of underlying language skills. Furthermore, interviews in practice depend on assessors' impressions which themselves require the combination of discrete items. True they are supposed to allow for spontaneity, but in view of my example of the neck-verse do they always do so?

To move from the Church and the law to that other great profession, medicine, let us consider the relation which we accept, most of us, in medicine between the doctor's diagnosis and the health we wish to have. Health (like proficiency) can be put on a scale so that we are all more or less healthy. However, we do not welcome or indeed expect the doctor who examines our heart or lungs or blood (whether when we are unwell or when we need an insurance policy) to use only direct tests of health, for example getting us to climb a mountain or run a marathon. What we want our doctor to do is to obtain in as precise a way as possible the best information about our body functions which s/he can relate to a health scale. In other words we are indifferent to the directness or indirectness of the medical tests used by a doctor to determine our body health. Why should an investigation into our language health be any different?

As far as proficiency scales themselves are concerned we cannot avoid the basic question which is just how valid they are. There is a sense in which all such scales are circular. The fact that they bring together proficiency level with authenticity is assumed to be itself an indication of value — but the question remains of just how valid the levels are. In the physical world we can indeed divide up nature in equal units and claim that the units are recursive, for example in measuring height, that each one is the same as the next. Just what is the status of the descriptors and of the example tasks at each level of the Proficiency Scales on offer. Are they predictors or are they criteria? How valid are they ?

What a test does is to predict a criterion. In my view the Proficiency Scales we have under discussion are a type of criterion. The interview is one such predictive test or test method facet, to use Bachman's terminology (Bachman 1990). There are many others and it would be wasteful and lacking in principle to the extent of over-enthusiasm if we were to restrict ourselves to a one type method. The testing literature provides suggestions of test methods and techniques, ranging in oral testing from discussion and interpreting to semi-direct tape tests of simulated oral performance (see for example Hughes 1989). ·

Finally I propose to make some comments on the Australian Second Language Proficiency Ratings Scale (ASLPR), itself an Australian version of the old FSI Scale.

First, its strengths. It is a positive virtue of the ASLPR that it focuses attention on the construct of proficiency. These levels, it implies, are the successive approximations, described in some detail that the learner makes as s/he approaches target, the native speaker goal of fluency. Experienced raters become so familiar with the meaning of the bands that they no longer need the descriptions and agree with acclaim on the placing of a candidate as a 1 or a 1+, even to the extent of sharing qualifications, a good 1+. (We are reminded of those -surely apocryphal-stories of examiners harmonising over beta double plus with just a touch of alpha in it). In this regard it is interesting to read Alderson's 1991 account of the development of the IELTS scales. He notes that even though IELTS was meant to be a fresh start after ELTS so much accumulated wisdom had been built up over the use of the ELTS bands in their ten

years of use that it was decided that the Revision Project...had to produce equivalent bands for the new tests' (Alderson 1991:83).

Now this is a very strong argument in support of the reliability of the ASLPR (and similar procedures). Trained raters are consistent, with one another and with themselves. That answers, at least in part the criticism made by Quinn and MacNamara of the ASLPR (1987:8), 'its built-in tendency to become a variable instrument'. Not wholly, of course, because as they point out there still remains the huge subjectivity of the interview as a means of eliciting the judgment data which trigger the rating given.

But it does not answer at all the question about validity. Being reliable/consistent is no guarantee of validity. The very strength of ASLPR, its security through consistency, its safe scaffolding, may persuade us into thinking that proficiency is now all safely tucked up in the ASLPR. That is the danger of over claiming. Nor does it resolves the doubt about measurement. In my book it is Pollitt's critique that hits home when he notes that a scale is not a measure, a point also made by Quinn and MacNamara (1987). This is in contradistinction to the exaggerated claims for these scales, for example by Burke: 'The ASLPR is an instrument which directly measures an individual's general proficiency in English in terms of his ability to carry out everyday, language specific tasks in real-life, non-specialist situations' (Burke 1983:2). Wrong! The ASLPR is not (in our sense) an instrument; it does not measure anything. It is not a direct test: take for example the descriptor for the S3 Minimum Vocational Proficiency :

> 'Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations on practical, social and vocational topics'.

In what sense is this a direct measure? Surely it would be more appropriate to call it an aide memoire! And in any case, even if it were a direct test, it is not logical to claim that a direct test measures general proficiency. What a direct test does is to test specific performance! That is the strong argument against the ASLPR, not against its helpful reminder to us that we should think in explicit terms about proficiency ,but against our gradually

allowing it to be used as if it were itself a measure, indeed in some contexts the only measure. It isn't and it should not be.

I incline towards Brindley's position (1989:10). Achievement and proficiency are too close to call. Alternatively, and for most purposes helpfully, we can argue that what we <u>always</u> measure is achievement. Our test instruments are <u>always</u> context sensitive. Achievement is <u>never</u> proficiency only an attempt to iconise proficiency. Proficiency scales are simulations, subjective, approximate and incomplete. We know only too well that tests and scores are unreliable and unstable ; we know that the equal interval scale is a myth (for example that the difference between a score of, say 2 and 3 is the same as the difference between, say the score of 3 and 4). But if that is true for test scores it is even truer for scales where to claim that the scale is anything more than nominal is sheer humbug.

The paradox is that through the attempt to refine proficiency scales by removing their defects (the imprecise and relativistic terminology — limited range, control of some structures, many error types) the precisioning of the descriptors tends more and more towards a list or bank of test items. Descriptors which are usable in an objective sense are test items. All the more reason for not making more precise, for acknowledging that a scale is not an instrument but a sort of metaphor to inform a judgement.

'Scales such as the ASLPR', says Pollitt, 'will, it seems to me, give little help to teachers or students since they do not describe the qualities of a performance. They are not criteria for good performance...they include no definition of what constitutes an acceptable level of performance in any task; they merely 'describe' a hypothetical set of tasks. I do not mean to say that such scales have no use in the planning of curricula and programmes of study, but they have no value to students or to teachers in formative assessment and teaching. They are not student-orientated' (Pollitt 1991:87-8). Such scales do not, says Pollitt, define minimum competence or minimum standards as they purport to do. At the same time he agrees with Alderson (op cit) that scale descriptors are addressed to several discrete audiences, including the test constructor, the test user, the learner.

It appears that proficiency scales can only tell us half the story. They are not and should not claim to be test instruments, ways of measuring. Assessment of learning needs the measure (the instrument) and the explanation (which may be in the form of a scale). Which is another way of saying that achievement and proficiency always need one another: achievement without proficiency is too local, too contingent; proficiency without achievement is unreal, unreliable and vague.

References

Alderson J.C. (1991) Bands and Scores. In Alderson J.C. and B. North (eds) *Language Testing in the 1990s*: 71-86

Alderson J.C. & B. North (eds) (1991) *Language Testing in the 1990s*. Modern English Publications in association with the British Council; Macmillan: London

Bachman L.F. (1990) *Fundamental Considerations in Language Testing*. Oxford University Press: London

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

Burke E.V. (1983) *Language proficiency : a review of the ASLPR and factors affecting proficiency development*. Unpublished paper prepared for the Department of Immigration and Ethnic Afairs, Canberra, January 1983.

Davies A. (1977) The construction of language tests. In Allen J.P.B. and A. Davies (eds) *Testing and Experimental Methods. Volume 4 of The Edinburgh Course in Applied Linguistics*. Oxford University Press: London: 38-104

Hughes A. (1989) *Testing for Language Teachers*. Cambridge University Press: Cambridge

Ingram D.E. (1982) *Report on the Formal Trialling of the Australian Second Language Proficiency Ratings (ASLPR)*. Australian Government Publishing Service: Canberra

Ingram D.E. (1990) Overview paper : towards the development of proficiency and other tests in Japanese as a foreign language in Australia' in Wylie E. (ed) *Assessment of Proficiency in Japanese as a Foreign Languge*. Conference Proceedings of an International Working Group sponsored by the Asian Studies Council, Canberra , June 1990. Asian Studies Council: Canberra:34-51

Pollitt A. (1991) Response to Charles Alderson's paper: Bands and scores. In Alderson and North (op cit): 87-94

Quinn T.J. & T.F. McNamara (1987) Review of Australian Second Language Proficiency Ratings. In Alderson J.C., K.J. Krahnke and C.W.Stansfield (eds) *Reviews of English Language Proficiency Tests*. TESOL: Washington DC: 7-9

Tests etc. referred to (year indicates first known application)

ASLPR : Australian Second Language Proficiency Ratings	1984
Cambridge : Certificate of Proficiency in English	1913
ELTS : English Language Testing Service	1980
IELTS : International English Language Testing System	1989
RSA : Royal Society of Arts Certificate in TEFLA	annual
TOEFL : Test of English as a Foreign Language	1965
VCE : Victorian Certificate of Education	annual