Introduction: Measures and Reports[1]

Alan Davies

1. All language learning is purposive, naturalistic second language acquisition (SLA) implicitly so, instructional SLA explicitly so. Instructional language learning is intentional and deliberate, that is to say it is predicated on learners' needs and expectations; it provides in its text-books, syllabuses, teachers' guides, band scales, tests and examinations operational definitions of what its purposes, its outcomes are. Whatever the quality of the instruction, therefore, descriptions of intended outcomes are always recoverable.

These descriptions seem to assume that the outcome (in the sense of 'ultimate attainment') should be either the native speaker (or some copy of the native speaker) or a defined proficiency. The first, the native speaker, seems more obvious, less abstract. The world, after all, is full of native speakers of the language under instruction. Proficiency, on the other hand, is less easily graspable, an abstract construct which we can only get at, so it appears, following Polonius in 'Hamlet', indirectly,

> *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)

Proficiency (unlike the native speaker) does not occur in nature, we have to invent it, define it, find something that stands for it.

---

[1]This paper is based in large part on Davies A. 1992 'Is language proficiency always achievement?' Melbourne Papers in Language Testing 1/1: 1-16

Melbourne Papers in Language Testing 1995 Volume 4.2 pp. 1-11. The Language Testing Research Centre, The University of Melbourne.

2. Valdman, introducing a special issue of Studies in Second Language Acquisition (10/2), refers to the so-called Proficiency Movement in the USA which

'represents an attempt to modify the nature of the foreign language curriculum in the direction of the acquisition of functional language skills' (Valdman 1988:121)

Valdman continues:

'There is scarcely any area of the field' of foreign language teaching 'in the US that has not been affected by' the 'attempt to institute a national metric based on demonstrated proficiency in the functional use of a foreign language and, more importantly, to define achievement in language instruction in terms of functional use rather than exposure to or command of a specific body of material.'

Valdman seems to imply that the Proficiency Movement notably through Rating Scales[2] may offer the elusive criterion definition/description. We might even surmise that this is a rejection of the native speaker type of target in favour of a proficiency defined target.

Proficiency, however, proves to be no less elusive than the native speaker. Debates about the nature of language proficiency have influenced the design of language tests and language testing research has been used in the validation of various models of language proficiency.

Proficiency may be defined in a number of ways[3]. For example:

---

[2]The American Council on the Teaching of Foreign Languages (ACTFL) and the Inter-Agency Language Round Table (ILR), formerly the Foreign Service Institute (FSI), Rating Scales, in particular the Oral Proficiency Test

[3]These definitions are taken from an entry written by Catherine Elder in a draft Language Testing Dictionary to be published by UCLES and Cambridge University Press

1) a general type of knowledge of or competence in the use of a language, regardless of how, where or under what conditions it has been acquired;

2) the ability to do something specific in the language, for example proficiency in English to study in higher education in the UK, proficiency to work as a foreign language teacher of a particular language in the United States, proficiency in Japanese to act as a tour guide in Australia.

3) performance as measured by a particular testing procedure. Some of these procedures are so widely used that levels of performance on them (for example 'superior', 'intermediate', 'novice' on the FSI scales) have become common currency in particular circles as indicators of language proficiency.

In its more portmanteau sense of general language ability, proficiency was widely used in the 1970s and early 1980s under the label general language proficiency, synonymous with unitary competence hypothesis. Proficiency has since come to be regarded as multifaceted, with recent models specifying the nature of its component parts and their relationship to one another. There is now considerable overlap between the notion of language proficiency and the term communicative competence.

One way of clarifying the notion of proficiency is to examine what it is not. To this end, 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 disentangle fully from the

circumstances of its use. On the one hand, apparently similar performance on as robust a test as the Test of English as a Foreign Language (TOEFL), or to a lesser extent the International English Language Testing System (IELTS) can be shown to vary in terms of factors such as mother tongue. On the other hand, on a proficiency scale such as the Inter Agency Language Round-Table (ILR) or the Australian Second Language Proficiency Ratings (ASLPR), the criteria influencing the bands allocated to different groups (eg a group of work place adults, groups of high school students or university postgraduates or foreign language/ international students) will not be identical. In other words, what count as criteria for a band in one context will not be the same as in another.

And for scales as for tests we cannot avoid the demands of validity. In its weak version validity emphasises the importance of the claim 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 ILR or the ASLPR. That is to say that the IELTS, for example, is said to be intended for academic purposes only; the ASLPR (and similar scales) seems to claim universal applicability. In so doing, they may 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). IELTS lays claim to only a very general coverage of academic English, reduced as it now is to two modules from three (itself a reduction from the 6 modules used for ELTS, the predecessor of IELTS). In comparison, the claims of ASLPR seem over-generous.

It may be suggested 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. However, as Alderson points out, when discussing the different audiences for scales, that creates its own problems. Commenting on the IELTS Band Scales for Reading, Alderson writes

'the production and public availability of band descriptors commit the test developer to a clearly untenable assertion' ie that future parallel tests 'measure "the same thing" in a highly specific way' (Alderson 1991: 76).

3. I turn now to a consideration of the relation between proficiency scales and proficiency 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' (Bachman 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 lack authenticity. They are always simulations of real life rather than real life itself. It is therefore the job of assessment not so much to replicate real life (because by definition that cannot be done) 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. If indeed that is the case, it is essential 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 involving proficiency scales typically uses the interview as a means of sampling oral language data which can then be related to the scale. Interviews may provide direct entry into the speaker's language ability, but they are also notoriously unreliable. It is true that there are ways of training raters and ways of pooling ratings but the extent to which assessments can be made reliable is a function of the amount of training and time and money that are available. In other words the more reliable the less practical. Direct assessment encounters (notably using scales in an attempt to approximate real life) are concerned above all with being (seen to be) valid. As such, 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, neither academically nor 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 rather than about the observed performance alone.

There is in direct tests such as 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. (Davies 1992). 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.

As far as the 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. But the status of the descriptors and of the example tasks at each level of the Proficiency Scales on offer is unclear unless they are a type of criterion.

4. In Australia the best known Ratings Scale is the ASLPR, based originally on the old FSI Scale. Its strengths are many. It is a positive virtue of the ASLPR that it focuses attention on the construct of proficiency. The levels on its scales are, it implies, the successive approximations, described in some detail, that the learner makes as s/he approaches target, the native speaker goal of fluency.

('there is nothing about the way I speak that suggests I am not a native speaker' ASLPR Level 5)

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 (for example) a 1 or a 1+,

even to the extent of sharing qualification, for example, 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 shown to be consistent with one another and with themselves. That answers, at least in part, the criticism made by Quinn and McNamara (1987: 8) of the ASLPR's 'built-in tendency to become a variable instrument'. It answers their criticism only in part because, as they point out, there still remains the huge subjectivity of the interview as a means of eliciting the judgement data which trigger the rating given.

It does not answer the question about validity. The very strength of the 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 Faustian danger of over-reaching. Nor does it resolve the doubt about measurement. Indeed, Pollitt (1991) expressly states that a scale is not a measure. This is in direct conflict with Burke who claims:

'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).

Such a claim is hard to reconcile with a straightforward interpretation of the descriptors, for example the descriptor for the ASLPR Speaking Level 3 (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.' (Can discuss own particular interests and special fields of competence with reasonable ease though some circumlocutions; vocabulary is broad enough that the*

learner rarely has to grope for a word and can readily overcome gaps with circumlocutions; accent may be obviously foreign; control of grammar good; able to convey meaning precisely in reasonably complex sentences or by using, with reasonable accuracy, a wide range of modification devices; fluency is rarely disrupted by hesitations; errors rarely interfere with understanding or disturb the native speaker; able to modify language to meet the differing register requirements of situations which are familiar in the learner's personal and vocational life but can make secure use of only high frequency collocations.)'

(Ingram 1982: 128)

It might be more appropriate to call such a descriptor an aide-memoire rather than a direct measure. Furthermore, even if the ASLPR were a direct measure it does not seem logical to claim that a direct test measures general proficiency. What a direct test does is to test specific performance. That is the really strong argument against the ASLPR (and similar scales), 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 shouldn't be.

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 the scale can only be nominal.

There are two sorts of difficult work in testing; one is the attempt to be explicit about kinds of ability, grades of performance; the other is the attempt to operationalise test meanings. There is no doubt in my mind that the latter is the harder and by far the more important of those tasks.

The paradox is that 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.

Here is Pollitt's opinion:

> *'Scales such as the ASLPR 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 on 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, he continues, define minimum competence or minimum standards as they purport to do.

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 both the measure (the instrument, the test) and the explanation or the report (which may be in the form of a scale). Best practice would require that they be used in conjunction.

5. Three of the four papers that follow were given at a special panel discussion on Scales and Tests held at the annual meeting of the Australian Association of Applied Linguistics in Canberra in September 1995. These papers, revised for this publication, were by Ingram, Hill and Scarino. Ingram provides a detailed discussion of the development of the ASLPR; as its principal author he is in a unique position to do so. Scarino discusses the use of scales in the teaching of languages other than English (LOTEs) in Australia, pointing to some of the serious difficulties she and her colleagues have encountered with scales in the assessment situation. Hill reports on a case study of a project which attempted a marriage between a testing and a scale approach to the development of a

Teacher Proficiency Test of Indonesian. Hill concludes that the test and the scale approach are indeed very different; if they are to operate well together, she opines, it is essential that they start from an agreed approach to assessment. North's paper was specially solicited for this publication. It represents two chapters of his recently completed thesis (North 1996). We include it because it provides a critical analysis of language proficiency scales which is both scholarly and wide-ranging. North is clearly an advocate of the use of scales but he is also aware of their inadequacies and excessive claims. His paper provides a salutary corrective, both to Ingram's very positive view of scales and to my own negative one.

6. References

Alderson J.C. (1991) 'Bands and Scores' in Alderson J.C. and B. North (eds) op cit: 71–86

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

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 Dept of Immigration and Ethnic Affairs Canberra

Davies A. (1992) 'Is language proficiency always achievement?' *Melbourne Papers in Language Testing* 1,1: 1–16

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

Ingram D. (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 Language* Conference proceedings of an international working group sponsored by the Asian Studies Council Canberra June 1990 Asian Studies Council Canberra: 34–51

North B. (1996) *The Development of a Common Framework Scale of Language Proficiency, based on a theory of measurement* PhD thesis Thames Valley University unpublished

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

Quinn T.J. and T.F. McNamara (1987) 'Review of Australian Second Language Proficiency Ratings' in Alderson, Krahnke and Stansfield (eds) op cit: 7–9

Valdman A. (1988) 'The assessment of foreign language oral proficiency' *SSLA* 10,2: 121–128