PASAA
Volume 28
December 1998

Assessment of Language Proficiency

D. E. Ingram
Professor of Applied Linguistics,
Director, Centre for Applied Linguistics and Languages,
Griffith University, Brisbane, Queensland, 4111, Australia.

I Importance of Proficiency

Proficiency is the single most important concept in language teaching and learning. Whatever else one aspires to do in a second or foreign language learning programme, proficiency or the ability to use the language for some purpose is always an important, generally the most important, aim. Even where it is not considered the central focus of the course, as it may not be, for example, in a traditional literature course, it will determine the extent to which the other aims can be realised.

II What it is: Definitions and Critical Features

In one of the most detailed recent discussions of language proficiency, the Education Committee of the Council for Cultural Cooperation of the Council of Europe in its *European Common Framework of Reference for Language Learning and Teaching* says that any form of language use and learning can be defined in the following terms:

> *Language use and learning are actions (among others) of a social agent who, as an individual, has and develops further a set of* ***general competences*** *and in particular* ***communicative language competence****; he or she draws on these competences in different kinds of* ***language activities*** *in order to process* ***text*** *(receptively or productively) in relation to specific* ***domains****, activating those* ***strategies*** *which seem most appropriate for carrying out the* ***tasks*** *to be accomplished. The contextualised use of general competences and in particular of communicative competence provides feedback which in turn leads to their modification.* [Education Committee 1996:8]

It goes on to say that *a social agent's* general competences *are the sum of knowledge, skills and characteristics .. which allow him or her*

to perform actions [Education Committee 1996: 8]. Thus, general language competence or language proficiency is the product of the knowledge, skills and characteristics related to language which allow a person to perform language actions or carry out language tasks. The *European Common Framework* goes on to analyse in great detail the nature of language ability or language proficiency and, in one part, speaks of "communicative language proficiency" as representing the observable features in a performance or *what happens when competence is put to use.* It differentiates between "competence" (the *underlying knowledge*), "performance" (*language use in context*), and "proficiency", which, it says, lies *between competence and performance* [cf. Education Committee 1996: 128].

Others have been more cynical about the notion of proficiency. Some fifteen years ago, Vollmer, for example, stated that:

> *...language proficiency is what language proficiency tests measure. This circular statement is about all one can firmly say when asked to define the concept of proficiency to date.* [Vollmer 1981: 152]

However, as the present writer commented in his paper *Assessing Proficiency* [Ingram 1985], this psychometric notion of language proficiency is analogous to the common definition of intelligence as what intelligence tests measure and all it does is to transfer the definition problem from *proficiency* to what we mean when we say that someone is more or less proficient than someone else or that X is better or worse at English or Thai or some other language than is Y. In fact, "proficiency" is an everyday, intuitive concept commonly used by people who speak a language, i.e., people know what they mean when they say that someone is better or worse than someone else in using a language, whether they can define it or not, they know what "proficiency" is and the task that language teachers, language testers or applied linguists have is to explicate what "proficiency" means and to measure it for the very real and practical purposes in real life for which people need to have available measures of language proficiency.

The present writer first started to grapple with the notion of "proficiency" and how to measure it when, in the course of his Ph.D. research in the mid-1970s, he wanted to describe the sorts of language skills students brought with them to university studies after five years of secondary school French and found that the results on matriculation examinations gave no indication of what students' actual or practical language skills were. Shortly after that, it became necessary also to try to specify the sorts of language abilities that learners had on entering and on exiting from adult migrant English programs in Australia and the sorts of skills that their courses should aim to develop. Similar needs were felt in the context of developing and assessing new "competency-based" foreign language programs for Queensland secondary schools. Out of these requirements, the present writer together with Elaine Wylie started to develop the *Australian Second Language Proficiency Ratings* in which language proficiency is characterised as encompassing

the tasks that learners can carry out and how they are carried out .

Definitions and descriptions of language proficiency differ greatly in their complexity. The International English Language Testing System, the IELTS Test, aims to measure proficiency in the context of candidates' ability to use language in academic and training contexts in English-speaking environments. The IELTS specifications speak of "language ability", discuss a number of "models of language ability" [cf. *The IELTS Specifications*, February 1996, Chapter 4]. These models and the *European Common Framework* referred to earlier examine the notion of language proficiency in great detail but, here, we shall limit ourselves to considering some common ideas that people have about language proficiency, i.e., we shall consider what we mean when we say that someone is more or less proficient in a language, that they "know" English, Thai or French and we shall consider some of the implications for assessment.

One traditional notion is that proficiency relates to knowledge, in particular, knowledge of grammatical rules and vocabulary and so, traditionally, tests of language proficiency or language ability have measured learners' knowledge of some features of language, characteristically syntax and vocabulary. This focus also commonly occurs in the sorts of tests that language teachers use in class. However, one can have a great deal of formal knowledge about a language without being able to use that knowledge fluently and automatically in the course of actual language performance and tests of formal knowledge of grammar or vocabulary do not correlate very closely with either real-life experience of practical language proficiency or with the results of tests intended to measure such proficiency [cf. Ingram 1985]. In other words, knowledge and proficiency are not the same thing and, although one cannot have proficiency without effective knowledge of the features that compose language, proficiency is more than such knowledge and also entails the ability to apply that knowledge in carrying out tasks in particular contexts and for particular purposes.

In contrast to the knowledge-based notion of language proficiency, a task-oriented approach to its definition and assessment regards knowledge as relevant only to the extent that that knowledge can be mobilised and applied in language tasks, in carrying out acts of communication. This accords with the real-life view that people are proficient in a language when they can do things using that language or can carry out communication tasks. Such a view is characteristic in task-oriented and functional and communicative syllabuses found in their most extreme form in the early "graded objectives" courses which focussed on identifying and presenting communication tasks while assessment focussed on whether learners could carry out those specified tasks. However, there are several limitations to such a notion, of which we shall consider just three. First, most communication tasks can be carried out in several different ways and at a range of different complexity levels from pointing to an object and uttering its name ("Butter" or "Butter, please" in a shop where the task is to purchase that object) to a complex utterance ("Could I have a kilo of butter, please,

preferably the unsalted variety"). Second, our intuitive notion of proficiency entails not only the carrying out of tasks but how they are carried out, whether the grammar is acceptable and native-like, whether the pronunciation used is standard, non-standard, like that of a native speaker of a recognisable variety, whether the content is organised and the argument developed appropriately to that culture, and so on. Third, use of a language entails not just being able to carry out tasks but carrying them out in ways that match what we regard as the requirements of different situations and their numerous contexts including where the utterances are made, to whom, by whom, and for what purpose.

Consequently, the ASLPR, referred to earlier, considers language proficiency not just to entail notions of the tasks that learners can carry out but also notions of how those tasks are carried out or the learner's total language behaviour. Starting with such a notion, language proficiency is described by referring to both language tasks and such features as the range of situations in which learners can operate, and the features of syntax, lexis, discourse, pronunciation, register sensitivity and flexibility, and cultural factors that can be observed. Learners' proficiency can then be assessed by eliciting examples of their maximum language behaviour which are matched against descriptions of language behaviour.

Approaches to defining, describing and assessing language proficiency that focus on total language behaviour encounter the serious problem of the complexity of language. Language is not just a single, unitary entity but consists of numerous components such as syntax, lexis, discourse, pronunciation and register that interact but also that may develop at different rates in different learning situations with different learners of different language and educational backgrounds. Thus, the notion of language proficiency has to allow for the fact that the different features of language may develop at different rates and that what ultimately matters is the total language behaviour that results. Thus behavioural approaches to defining, describing and measuring language proficiency aim to provide a global picture of the behaviour observable at each level rather than provide a checklist of the features mandatorily to be observed at any level. In any case, as the *European Common Framework* observes, the aim of proficiency descriptors is to provide an overview of language performance and language behaviour and so proficiency descriptors are necessarily to be considered holistic rather than as focussing in on variable descriptive detail [cf. Education Committee 1996: 130]. For assessment purposes, this approach seems to entail a subjectivity that seems not to be present if one can list and assess specific features of grammar or lexis but that degree of subjectivity is necessitated by the sheer complexity and redundancy of language: while it is "neat" to be able to assess specific features independently and objectively, it is not the nature of language to be neat but to be complex with interacting components whose interaction cannot be simplistically calculated.

It is also necessary to distinguish a person's underlying proficiency from its realisation in any particular task in any particular situation and scales such as the ASLPR speak of the

sorts of tasks rather than the actual tasks that learners can carry out. The reason for this is that one's ability to carry out a particular task in a particular situation depends not only on one's proficiency but also on one's familiarity with the situation. Thus, for example, someone who has never used public transport may have some difficulty in mobilising the language readily or appropriately to buy a ticket, use a timetable to find out bus numbers, routes and times, or to ask for appropriate information even though he or she has mastered such functions as "seeking information" in other contexts. As another example, someone who has not previously encountered the task of leaving written instructions for a door-to-door vendor may be unable to write an appropriate note to change a daily order even though in some other situation (e.g., writing a note for a teacher to explain an absence) they are able to perceive a relationship and the information to be conveyed and to select from their language repertoire the minimum language forms needed to carry out the task. In other words, one has to distinguish a person's underlying language proficiency from that person's ability to carry out an absolutely specified task in a specified situation. Clearly this distinction has important implications for the design of language proficiency assessment instruments and throws into doubt the validity and reliability of pre-determined sets of items that seem to limit the notion of proficiency to performance on that set of tasks in that set of situations. For this reason, the assessment approach characteristically used with the ASLPR and some other proficiency rating scales gives the interviewer freedom to vary the actual tasks so as to distinguish learners' underlying proficiency from their ability to perform specific tasks in specific situations.

The notion of general proficiency as the underlying proficiency that is generalisable across situations is an important issue when one tries to relate the notions of general proficiency and language competencies. In recent times, especially in Britain and Australia, the specification of, teaching towards, and assessment of vocational competencies has become a major issue that is influencing all aspects of education and training. Vocational language competencies focus very precisely on the tasks that a learner can carry out and, in particular, the vocationally relevant tasks, without asking whether the ability to perform them generalises across the language and to other situations or not. Thus vocational language competencies refer to the use of the language in particular vocationally relevant tasks to carry out jobs required by particular vocations. They are exemplified by the British National Language Standards or the National Reporting System in Australia which identify specific vocational language tasks without reference to their generalisability or their developmental complexity.

Messick [1994] makes a distinction between performance tests that are construct-centred and those that are more specific in orientation and are task-centred. Scales such as the ASLPR claim to measure the underlying ability that is manifested as the learner seeks to carry out various language tasks. The inevitable question arises as to whether one can generalise from how a learner carries out any particular task to that underlying ability;

instruments such as the ASLPR insist that if one is seeking to measure underlying ability it is essential to use more than one task. In a task-centred approach, as found in the specification and assessment vocational competencies, generalisability is not so significant.

Messick states that, in construct-centred assessments where, for example, one is attempting to measure language proficiency, one should determine what complex of knowledge, skills and attitudes should be assessed and then determine the behaviours or performances that should reveal these constructs. The nature of the constructs guides the selection and construction of tasks as well as the scoring criteria. On the other hand, in the task-centred approach or competency approach, it is necessary to determine the actual performances that we want students to be good at and then decide what tasks will elicit those performances. Messick states that this approach is most suitable in those fields where the mode of teaching emphasises repeated demonstration, practice and critique. Although he acknowledges that there are arguments on both sides, he comes out in favour of the first alternative, seeming to favour the underlying attribute approach, stating that

> *Principles like relevant knowledge and skills, rather than domain related tasks and performances ought to drive the development, scoring and interpretation of performance assessment.* [Messick 1994: 16]

The possibility of different rates of development in different components of language and the difference between underlying proficiency and its realisation in particular tasks are also important issues when it comes to considering the different macroskills of speaking, listening, reading and writing. It is a common shorthand to speak of a person's level of language proficiency as though that level is the same for all four macroskills. Thus, for example, TOEFL provides a single numerical score such as 575 unless one specifically asks for additional Speaking and Writing scores. The IELTS provides an overall bandscore such as 6.5 or 7 and most secondary school result sheets provide a single figure such as 75%, High Achievement, Credit, and so on. However, most language teachers are conscious of the fact that different learners in their classes may be better in one macroskill than another and, in fact, the empirical evidence against any unitary notion of language proficiency is strong [see Ingram 1985]. The implications of this for language proficiency assessment are important since it implies that any test that provides only a single measure of language proficiency will give, at best, only a very general or partial indication of what the learner can do with the language. Consequently, instruments such as the Australian Second Language Proficiency Ratings, the ACCESS test of English taken by applicants for migration to Australia, and the IELTS test all provide a profile of the learner's proficiency in some form such as, for the ASLPR, S:3, L:3+, W:3+, R:4.

Of the many other aspects of language proficiency that are crucial to its assessment, just one other will be considered now, one that

is especially crucial to the definition of language proficiency. Language proficiency refers to that underlying ability that is realised in actual performance in actual tasks in actual situations. It does not equate to the ability to communicate and even though, in language teaching, we may ultimately be concerned with developing learners' ability to communicate, it is very questionable as to whether, in language testing, we can or should focus on that or that is a valid and reliable instrument to measure the ability to communicate or what is sometimes loosely termed "communicative competence"[1] could be developed. A person's ability to communicate depends on many factors that lie outside language proficiency, factors such as personality, introversion and extroversion, intelligence, educational background, and the willingness of an interlocutor to accommodate a learner's non-native forms. Bachman tries to accommodate for some of these additional features of communication in his model of language ability [e.g., Bachman 1990; Bachman and Palmer 1996] though some of them clearly entail features that go well beyond language itself and even beyond the language learner. In other words, a statement about a person's language proficiency does not necessarily indicate that person's ability to operate effectively in particular situations, for example, to engage in social interaction at a cocktail party, to carry out particular work, or to succeed in an academic program, all of which entail things that go well beyond language itself. The problem was vividly expressed some twenty years ago by Sollenberger when he stated:

> *The person's so-called language proficiency, while it may have been quite accurate in technical skill terms, did not mean effectiveness in* ***communication****. In some cases, it may have enabled the person to misrepresent or foul up more effectively...I'm sure we all know people who talk nonsense fluently.*
>
> *On the other hand I know people who butcher the language, whose accents are atrocious and whose vocabularies are limited. For these reasons we give them a low proficiency rating. Yet, for some reason, some of them are effective communicators.* [Sollenberger 1978: 8]

In summary thus far, the present writer's approach to language proficiency as reflected in the Australian Second Language Proficiency Ratings, is to see proficiency in a second or foreign language as the learners' practical language ability, to entail the sorts of tasks they can carry out and how they carry them out, and to be manifested in the learners' language behaviour. Language is to some extent situation-dependent though there are features that occur more or less widely across the language in many situations. General proficiency (in contrast to specific purpose proficiency) is seen as referring to those abilities that generalise across a variety of common, everyday situations, it relates to ability in those commonly occurring features and the tasks they realise, that occur across many situations, characteristically but not exclusively in situations of everyday life that

[1] Not "communicative competence" in the Hymesian sense which is close to the way in which "proficiency" is used in this paper.

most people commonly encounter in the course of everyday human existence. Specific purpose proficiency, on the other hand, refers to ability in a more limited range of the language where the language is used for particular purposes in particular situations. Thus one can speak of proficiency in the language of civil engineering, French for academic purposes, and so on. Proficiency (whether general proficiency or specific purpose proficiency) also contrasts with the concept of language competencies, characteristically used in the context of vocational language competencies.

Consideration of the sorts of issues discussed in trying to define language proficiency has profound implications for how one describes and assesses it. First, we shall consider some alternative approaches to assessing language proficiency and then consider the nature and development of scales that are used to describe and assess proficiency.

III Some Alternative Approaches to Assessing Language Proficiency

The *European Common Framework* [Education Committee 1996: 136 - 143] discusses 13 pairs of contrasting approaches to proficiency assessment:

1	Achievement Assessment	Proficiency Assessment
2	Norm-referencing (NR)	Criterion-referencing (CR)
3	Mastery Learning CR	Continuum CR
4	Continuous Assessment	Fixed Assessment Points
5	Formative Assessment	Summative Assessment
6	Direct Assessment	Indirect Assessment
7	Performance Assessment	Knowledge Assessment
8	Subjective Assessment	Objective Assessment
9	Checklist rating	Performance Rating
10	Impression	Guided Judgement
11	Holistic Assessment	Analytic Assessment
12	Series Assessment	Category Assessment
13	Assessment by others	Self Assessment

[Education Committee 1996: 137]

In fact, one can classify approaches to language testing and proficiency assessment in particular in many different ways, each "cut" through the field highlighting different contrasts. Using an historical "cut", changes in language testing can be seen to have reflected changes in our understanding of the nature of language and language learning. When language learning was seen as a process of learning grammatical rules and vocabulary and then "rewriting" from one language to another, language proficiency was measured by tests of grammatical knowledge and translation. When, in the days of behaviourist psychology

and structuralist linguistics, language was seen as a set of patterns learned by stimulus-response habit formation, tests focused on the individual patterns or "discrete points" in, typically, multiple choice tests of elements of the language. Later, when the complex, integrated and highly redundant nature of language was noted, language tests emerged that used the principle of redundancy, deleting items and assessing the extent to which learners could replace them using the redundant features of the text to identify what was deleted. In the late 1970s and throughout the 1980s, the communicative nature of language came to the fore and so language tests and language proficiency assessment approaches came to focus on learners' ability to communicate and ranged in form from approaches focussing on the discrete tasks learners could carry out through to more complex approaches that focus on the learners' total language behaviour as they use the language for normal communication purposes. The last approach has often included the use of scales that describe how language behaviour develops and which are used either to explicate the results on other types of tests or are themselves used as the learners' language behaviour is observed and matched against the scale descriptors. Other useful "cuts" include such contrasts as the following:

developmental vs. non-developmental: i.e., tests that relate to some notion of how language develops in contrast to tests that simply assess performance irrespective of any developmental sequence against which proficiency development might be measured.

discrete point vs. integrative and behavioural: i.e., tests that assess whether learners have mastered discrete points or features of, for example, grammar, vocabulary or functions in contrast to tests such as cloze or dictation that try to integrate the assessment within a total language act or focus on total language behaviour.

norm-referenced vs. criterion-referenced: i.e., tests that essentially rank-order learners on the basis of their scores in contrast to tests that match the learners' performance against specified criteria.

intrusive vs. non-intrusive: i.e., tests commonly "intrude" on learners' activity by requiring them to take some form of formal assessment that is different from their real-life use of the language but one might also assess proficiency by simply observing the learners' use of the language in the course of their normal activity.

non-adaptive vs. adaptive: i.e., a test may be pre-determined in all of its features but, as was noted earlier, language is partially situation dependent and if learners happen not to have encountered a particular task or a particular situation, they may be unable to perform to their maximum ability. In contrast, an adaptive test, leaves it open to the assessor to vary the language tasks that the learners are required to perform so as to adapt to, for example, the learners' stage of development and

explore what they can do rather than what they cannot do.

One of the more useful "cuts" is that between **indirect**, **semi-direct** and **direct** tests, the directness relating to the extent to which the assessment procedure focuses on the learners' language proficiency or their actual language behaviour. **Indirect tests** essentially test one thing, characteristically knowledge of grammar or vocabulary, and try to say something about something else, in this case, proficiency (which, as we have seen, includes elements of such knowledge but also other things such as the ability to apply that knowledge in carrying out language tasks). The relationship is made between the test results and proficiency by, usually, psychometric or norm-referencing procedures in which the results are distributed over a normal distribution curve and then cut-off points are identified for different proficiency levels either on the basis of "gut-feeling" or, more desirably, by comparing the different learners' scores with what they can do in other contexts or on other tests. Some approaches try to match scores on indirect tests with proficiency scale ratings or, as with a test like TOEFL for example, with learners' subsequent performance in academic contexts. Item response theory and modern statistical procedures have also led to attempts to relate actual items closely to levels on proficiency scales though at this stage such close equating of items and proficiency levels is fraught with danger because of the many factors that determine the difficulty level of items. Typical of indirect tests are discrete-point tests, sometimes called "analytic" or, incorrectly, "objective" tests, in which language knowledge and language behaviour are analysed into the smallest possible units and knowledge of or ability to use those units is assessed. Two major problems exist with indirect tests: first, language performance and hence language proficiency are more than the sum of a multitude of discrete bits, in real-life language use, language items don't operate separately but together, supporting each other in meaning and dependent on each other structurally, and part of the skill of language use involves being able to put all the items together and to comprehend them when received together. Secondly, the interpretation of the results on indirect tests is also fraught with difficulty. Thus, for instance, a score such as 4 out of 7, 80% or 525 says nothing about the level of the learners' practical skills or what they can do unless such scores can be related to performance scales in which actual language behaviour is described.

Semi-direct tests, which, as was noted earlier, are one category of integrative tests, include such itemtypes as cloze, dictation, white noise and interlinear tests and seek to integrate the language components into a total language event and to test knowledge of them or the ability to use them in that total event. They effectively are intended to reflect the extent of the learners' language proficiency development by measuring the extent to which they are able to use the language's inherent redundancy to replace the missing item. Though such tests resemble indirect tests in that they seem to be testing discrete items, scores are processed psychometrically and can be related to proficiency estimates in the same way as indirect tests are but, in addition, the fact that a total language event is used puts these tests closer to real language performance

or the demonstration of real language proficiency. In the latter context, they resemble direct tests.

Direct tests are also "integrative" and focus on actual language behaviour, i.e., they characteristically are used to measure proficiency by having learners perform actual communication tasks while their language behaviour is observed and rated against proficiency descriptors that form a scale. Since the assessment procedure and the scale descriptors themselves focus directly on what learners can do and on their language behaviour, the assessment is more direct, interpretation of the test performance and scores is much easier, and the results are more immediately and practically useful. They are not the only way in which proficiency can be measured and there are many contexts in which their use is difficult but many proponents of proficiency in language education (including the present writer) would assert that direct tests are the most effective approach.

IV Scales

The interest that now exists in proficiency and its assessment has led to strong interest in the use of scales for direct assessment purposes, as a way of more concretely interpreting scores on other assessment types, and for their contribution to curriculum development. Not the least of their advantages is that they provide a common language by which to talk about the otherwise nebulous concept of language proficiency. Thus, for instance, the Australian Second Language Proficiency Ratings are characteristically applied in an interview situation in which learners' maximum language proficiency is elicited, observed and matched against the scale's descriptors ranging from zero to native-like through a total of nine described levels and three undescribed levels. The ASLPR is also used to interpret more traditional test types and to provide a framework within which tests and language curricula can be developed. In concretising and providing a common language about proficiency, scales such as the ASLPR are also used for many purposes in everyday life from cataloguing easy reading materials to identifying language proficiency levels in legislation or for employment.

Since scales play such a major role in the assessment of proficiency, we shall now consider their nature and how they are developed.

IV.1 *All Scales and Tests are Compromises:*

Scales, like all assessment-related instruments, are compromises designed for certain purposes and for use in certain contexts, those contexts including such variables as who is administering them, in what sort of situation, with what sort of candidates, under what quality control conditions, and for what end-users of the test results. In that sense, one cannot say that scales are better or worse than other sorts of assessment-related instruments but only that in such and such a context for such and such a purpose they may be more or less appropriate.

All tests and scales are also compromises in the extent to which they represent real life

language performance. A test situation is itself a part of "real life" but the extent to which any assessment process can be said to reflect real life depends on the purpose of the assessment, the context within which the assessment results are intended to be used, and the extent to which the items or activities used to elicit language performance stimulate in the candidate language processes that are similar to those that occur in the context within which the assessment results will be used. So, for example, an essay-writing item may be quite inappropriate in a test of general proficiency since it could be said not to match the sort of writing one does in everyday life but it would probably be entirely appropriate in an Academic Purposes test. The assessment situation itself imposes constraints that necessitate more or less compromise with real life language performance but the art of the tester is, within the context and purpose of the assessment, to approximate as closely as possible to activities that are relevant to the real-life context within which the test results are to be used.

Again, with scales, evaluation of the extent to which they are descriptions of "real life" will depend on their purpose and the context in which they are to be used. A scale such as the Australian Second Language Proficiency Ratings (ASLPR) [Ingram and Wylie 1979/1985 and Wylie and Ingram 1995] seeks to describe language proficiency as it develops from zero to native-like and is used to assess practical language skills. As such, it aims to describe real-life language performance but it is necessarily selective and suggestive of real-life language behaviour rather than replicating it in all its multitudinous features, which, in any case, differ in every situation where language occurs. Again, a scale is necessarily a compromise made between, on the one hand, descriptions of real life language behaviour and how that behaviour develops and, on the other hand, what the authors assess to be manageable by the users of the scale. The complexity of language and its variations from situation to situation according to who is using it, to whom, in what medium, in what location, for what purposes, and about what topics, all mean that, if the scale descriptors actually attempted to match real life language performance, they would be unmanageable: they are, perforce, a selection that reflect rather than match. In addition, as verbal, written descriptions rather than videotapes of language use, descriptors are already moving away from real life though adequate training programs in the use of the scale make much use of recordings to assist users of the scale to relate the verbal descriptions to real-life language behaviour. In addition again, to make the scale and its descriptors generalisable across the possible instances of general language behaviour, the scale seeks to describe underlying language behaviour, to prompt in the user an image of the underlying behaviour that is realised in each real life instance of language use. In brief, scales are not "real life" but they seek to suggest real life language behaviour and are validated against real life language behaviour or performance.

In considering scales, one has always to distinguish between the scale itself and how it is used. Scales such as the ASLPR, ACTFL, FSI/ILR, or the speaking and writing scales of IELTS or ACCESS are used by being matched against language performance, typically in a

live or simulated interview situation. As with other tests, the extent to which the interview matches real life language use depends on the total context within which the assessment is being administered and the compromises that have to be made to accommodate the test-taking situation. The ACCESS oral interaction interview departs considerably from real life, largely because it was felt by the developers that it should be parallel to the simulated oral interview required for use in situations where it is not possible to use live interviewers and administered in a language laboratory using recorded prompts to elicit candidate responses for recording on tape and subsequent assessment. In contrast, the elicitation procedure in a standard ASLPR assessment seeks to elicit language that approximates as closely as possible to that which occurs in real life language performance, it is adaptive to the candidate's needs and level, and it seeks to reduce the intrusion of test method that occurs in less adaptive tests and in tests where the context requires a greater compromise with real life. However, again, the practical constraints of assessment and the need to elicit maximum language performance for observation and rating necessitate some compromise with real life (as does the very fact that all parties in a formal interview know it is an assessment situation rather than some other aspect of real life). Thus, the "live" oral assessment interview is a compromise with real life though one whose method is designed to intrude less on the reality of candidates' performance than in some other approaches to assessment.

One possible way of assessing whether scale descriptions approximate to real life language behaviour is to consider whether language learners and users find that they match with their own assessments of their own language behaviour. Thus self-assessment, with all the qualifications that this approach requires for high stakes testing, may show whether learners believe that the scale descriptors match what they know about their own language behaviour. Studies with the self-assessment versions of the ASLPR show a high correlation between self-assessment, real life observations, and formal interview-based assessment with trained raters using the ASLPR and suggest that the compromises that are made in selecting the content of the descriptors, in sequencing the scale descriptors, and in applying them to assessment purposes are defensible.

IV.2 *The Nature of Scales*

Scales may take a variety of forms but undoubtedly the commonest form is a graduated series of descriptions of language behaviour or of selected aspects of it. The *Shorter Oxford English Dictionary* definition is relevant in offering as the definition of a scale:

> *I. a ladder...III. a set or series of graduations (along a straight line or curve) used for measuring distances, etc.* [Onions 1967:1798]

The ASLPR, for example, seeks to describe the changes that are observable in language behaviour as a learner's proficiency develops from zero to native-like, i.e., from inability to use the language for any practical purpose to an ability indistinguishable from that of a

native speaker of the language. The behaviour that is described is characterised as encompassing the sorts of language tasks that learners can carry out and how they are carried out, i.e., the linguistic forms that are used in carrying out those tasks.

The ASLPR seeks to provide a fairly comprehensive picture of language behaviour related to its view of how interlanguage develops. Other scales take quite different forms. The IELTS Bandscales and the ACCESS scales provide very concise descriptions that do little more than suggest the sorts of language behaviour that can be observed at any level. Other instruments, like many vocational competency specifications, have some notion of increasing complexity in the tasks that can be carried out at different levels but the specifications are not so much related to some notion of how language develops as to the developers' own notions of what are more or less complex tasks in the workplace, considering not just language but also the other skills the particular workplace requires. The recently released *National Reporting System* [Coates *et al* 1995] in Australia seems to take this form in the context of specifying language and literacy competencies.

Several different types of proficiency scales may be identified according to how they are constructed, what they attempt to measure, and what criteria they contain [cf. Ingram and Wylie 1991]. Without pursuing all of the issues in detail, the contrasts identified include:

- **Whole vs. Part**: Proficiency scales may relate to the whole span of proficiency development or only a part of it.

- **Serial vs. Threshold**: Scales may provide a series of intermediate points between two levels or a threshold level may be described with more cursory (if any) descriptions of behaviour above and below that level.

- **General vs. Specific purpose**: A scale may aspire to describe general proficiency or proficiency in some specified area of the language.

- **Task-only vs. Total or underlying behaviour:** Scales may seek to describe in some way total behaviour or underlying behaviour that is generalisable to a variety of contexts or they may select only certain specific tasks which are graduated in some way along a scale, as is exemplified in some vocational competency specifications (as discussed earlier).

- **Proficiency vs. Course achievement:** Scales may seek to describe general proficiency at each level or may seek to provide performance descriptions that are related to specific course content and so constitute course achievement-related scales rather than general proficiency scales. Graded objectives of the sort popular in the late 1970s and early 1980s provide one extreme example of this approach.

- **Macroskill-specific vs. Overall:** Some scales, such as the ASLPR, describe language behaviour in the macroskills separately. Others combine two or more macroskills into one scale, as, for example, was the intention in the original oral interaction scale for the ACCESS test which sought to combine listening and speaking into the one scale related to conversational use of the language. While the IELTS provides for macroskill-specific ratings, it also provides an overall bandscale which seeks to describe general language behaviour in a way that is intended to relate to all of the macroskills.

- **Absolute vs. Global:** Some scales such as the FSI Scale in the 1970s are absolute in the sense that all criteria within each descriptor must be fulfilled before a learner is rated at that level whereas other scales, including the ASLPR, try to recognise the complexity of language behaviour and its development. As discussed earlier, such scales aim to provide a global picture of language behaviour, they allow for the fact that some of the parameters of change may develop more or less rapidly within the total skein of language behaviour, and they assert that it is the total picture of language behaviour that is more practically significant than the fact that certain parameters may be more or less developed than the characteristic development pattern would suggest at the particular level in question.

- **Analytic vs. Holistic:** Some scales seek to provide quite detailed statements about, for instance, specific aspects of grammatical development at particular levels on the scale. Others make more generalised statements about general development in, for instance, grammar. The ASLPR attempts a compromise between these by making generalised statements in the General Description column but giving specific examples in the middle, example column in order to concretise the general description and to assist the readers to interpret the general descriptions.

- **Empirical vs. Washback effect:** Probably the majority of proficiency scales seek to describe language behaviour as it is observed and so claim to be, in some sense, empirical. Others may be at least as much concerned with the washback effect of the scale on teaching and so seek to describe what is considered to be desirable or to be desirable elements of behaviour to aim at in a course.

IV.3 *Developing Scales:*

The validity of scales is closely dependent, amongst other things, on how they have been developed. Different approaches are possible but here reference will be made to just two examples, that by which competencies may be specified (whether in scalar form or not) and the approach adopted in the ASLPR.

In specifying competencies, tasks are identified that the learner or worker is expected to be able to carry out. The focus is not on how they are developed but on the target tasks or the tasks that constitute the activities of, for example, a particular vocation. For that reason, they would seem, intrinsically, to be of less interest to teachers charged with the task of developing learners' language than to employers and recruitment officers charged with the task of recruiting people able to carry out specified duties. Competencies are thus specific, not necessarily related or representing underlying abilities, and static in time rather than developmental. They are developed by observing, for instance, vocational activity, and analysing and specifying its task elements. In contrast, a second or foreign language proficiency scale which seeks to describe how language develops across a span (typically zero to native-like) is not only concerned to identify what a learner can do in the language but to sequence those specifications in some rational and empirical manner related to how a second or foreign language develops.

Ideally one would use the detailed findings of developmental psycholinguistics to develop a proficiency scale but, in practice, psycholinguistics has tended to be fairly minutely focussed on elements of the language rather than on the total context within which language is used. Scale development need be no less empirical even though scales such as the ASLPR also attempt to capture native speaker intuitions about their language and its development and make these intuitions concrete by capturing them in descriptions of observable language features.

The development of the ASLPR (an activity which has proceeded now for some 18 years since 1978) can serve to exemplify how scales may be developed. The following actions were taken:

1. A notion of proficiency was adopted and evolved as the scale developed: proficiency was defined as the ability to use language for purposes of communication, the notion encompassing both the kinds of communication tasks that learners can perform using the language and the kinds of language they use when performing those tasks.

2. Drawing on the intuitions and experience of the authors and others (including the authors of other scales), Ingram and Wylie sketched descriptions of language behaviour and how it develops. Some scale developers elaborate on this step, eliciting "indicators" from many teachers and other participants, and calibrating them to arrive at descriptors that represent common agreement on the elements of language behaviour observable at each level.

3. The initial descriptors were then tested out, elaborated and refined in interviews with learners throughout the proficiency span. The aim of these interviews was to elicit the features of the learners' language so as to evaluate whether the descriptors were comprehensive, coherent and consistent. This process has continued

over the years since the scale was first developed with the same process occurring with the basic scale and with different versions so that the newest general proficiency version released in 1995 is the product of empirical studies involving many thousands of learners of English and other languages, including their use in specified purpose contexts.

4. At the same time, the emerging scale was compared with evidence from psycholinguistics to assess whether it was compatible with those general findings. Ideally, one would like also to have the detail of the general descriptors and the specific examples of language tasks and language forms evaluated against developmental studies and the detail of psycholinguists' findings used to modify (if necessary) and elaborate the general description column and the specific examples. At present, however, the comparison has been in terms of the general compatibility of the scale with the theories and findings of developmental psycholinguistics rather than with the detail.

5. The scale was formally trialled using adult and adolescent learners, especially of English but also of other languages [see Ingram 1985]. This formal trialling essentially assumed that, if the series of descriptors making up the scale really did reflect second or foreign language development, if they described features of the language that generally do co-occur, and if they were comprehensible and manageable, teachers trained to use the scale would be able to interpret the descriptors consistently and apply them reliably.

6. More recently, statistical processing has been used in other ways to check the scale's validity. In recent years, it has become possible to apply such analyses as Many-Faceted Rasch Analysis to ASLPR data collected in the course of normal use of the scale in the Centre for Applied Linguistics and Languages at Griffith University and, with these and other techniques, to assess not only the validity and reliability of the assessment procedures but also to assess the adequacy of the scale itself. In one recent study, Tony Lee analysed the results of more than 300 assessments on each of the four macroskills with the aim of establishing whether the levels in the scale actually do represent a progression from zero to 5 along a common dimension, whether the four macroskills do form a reliable measurement variable, and whether the ordering can be the basis for construct validity [see Lee 1993]. In summary, Lee concluded:

    1. The ordinal nature of the ASLPR levels is established.

    2. The nature of the four macroskills as sub-scales of the ASLPR scale is established.

    3. Using data gathered over a two year period, the ASLPR scale

and its sub-scales seem able to uncover a proficiency development path for ESL learners from diverse L1 backgrounds and age groups.

6. There is little noise in the system. (In practice, this meant that, of the more than 300 candidates in the study, each assessed on four macroskills, only one rating for one macroskill was identified by the program as misfitting.)

It is evident, however, that not all scales have adopted such a long and detailed process for their development and their on-going re-development and validation as has the ASLPR. Undoubtedly some (even some of considerable international significance) have apparently been written more or less off the tops of their authors' heads with little if any subsequent validation. One has to question whether that should be the case where ratings assigned against a scale have significant bearing on candidates' lives and opportunities.

IV.4 *Validity and Reliability:*

Many of the issues already discussed, especially in relation to the development of scales, touch on issues of validity and reliability that can be discussed only briefly here. Some scales amount to no more than a statement of the intuitions of their authors. Others are non-developmental listings of competencies while a scale such as the ASLPR aims to be more comprehensive and more empirically based in describing underlying language behaviour and sequencing it developmentally. The validity of intuitions is, at best, difficult to assess or reject but the validity of scales such as the ASLPR is assessable against such criteria as their underlying construct, their comprehensiveness, their consistency, their coherence, the adequacy of their description of real life language behaviour, and the findings of developmental psycholinguistics. However, as has already been noted, the validity of a scale can be assessed only against what is known about what it seeks to describe, its purpose, the context in which it is to be used, and the context within which the assessments are to be used. Predictive validity studies would also seem to be useful in assessing the extent to which scales and assessments using them reflect real life language performance but, in reality, such studies are difficult to structure convincingly. However, recent studies by Kellett and Cumming [1995] looking at student performance in subsequent TAFE vocational courses and by Weston [1995] looking at students in university programmes found a strong correlation between ASLPR levels on entry and success in their subsequent courses.

The reliability of scales seems to hinge on the extent to which different users can interpret them in the same way. Some reference has been made to this issue in discussing the development of scales. In many instances, of course, it is not just the reliability of the scale that is in question but the reliability of the assessment procedures that are used to rate learners against the scale. The nature of those assessment procedures, whether interviews, simulated interviews, pencil-and paper tests or some other approach determines the nature of

the reliability studies to be conducted. Reference was made earlier to some of the studies conducted to assess the reliability of the ASLPR and of the interview procedures commonly associated with it. In the case of short simply worded scales, reliability would seem likely to be lacking because so much of language behaviour "falls between the cracks" and, not being described, it leaves the scale user to guess at where to locate other observed features of language behaviour. Such scales in, for instance, ACCESS and IELTS listening and reading are not used for direct assessment but for normative distribution and interpretation of other test results giving, at first glance, possibly higher levels of reliability but more questionable validity. On the other hand, more elaborated scales such as the ASLPR, ACTFL or FSI/ILR seek to be more comprehensive in their descriptors but are also more complex to use and interpret and, with these as with all scales, users still have to make the match between the descriptions provided and the real life language behaviour they seek to represent. Consequently, the authors of the ASLPR consider that training in the scale and its use is essential and that users need the opportunity at frequent intervals to, in a sense, re-validate their interpretation of the scale in training sessions in which they have the opportunity to re-consider the scale systematically, observe learners, interpret the scale, and check their own assessments against others'. As the data in the Lee study cited earlier demonstrates for the ASLPR, where this occurs, this approach to proficiency assessment can yield high levels of reliability. Consistently, data on user ratings collected at the end of short, "advanced" ASLPR training sessions conducted by Elaine Wylie have yielded test-retest reliability figures in the range of .87 to .89 with higher figures where the assessments are conducted by well-trained and supervised assessors with extensive experience in interpreting the scale and conducting interviews.

V CONCLUSION

This paper has sought to provide a brief overview of some of the factors to be considered in the assessment of language proficiency. There are many other issues that warrant consideration but which have been mentioned only in passing or not mentioned at all. Some of these include, for instance, cultural appropriacy, how proficiency ratings can be related to more traditional statements of school grades, the useful distinction for language teachers between tests of course achievement and general proficiency assessments, the relevance of proficiency to summative and formative assessment, the usefulness of self-assessment scales, and the application of proficiency and its assessment in many real-life contexts. The irony of language proficiency and its assessment is that language proficiency, interpreted as a learners' practical language skills or their ability to use the language for real-life communication purposes, is what learners and the general public tend to think that language learning is all about but, in second and foreign language education, we have traditionally been obsessed over the teaching and learning of the detailed features of the language to the point where we often lost sight of the larger picture of what learners could use the language for. Language does, of course, consist of a multitude of discrete and inter-related features that must be taught and

learned or there can be no substance in the language and no flexibility in its use but the notion of proficiency has enabled language educators also to re-focus on the broader picture and the practical value of what they are seeking to do. Not least, the notion of language proficiency has forced language teachers and language testers to re-evaluate how they go about assessing language learning and to re-assess the relevance of their assessments to learners' practical language skills and their desire to use the language for real-life communication purposes.

The Author

David Ingram is Professor of Applied Linguistics and Director of the Centre for Applied Linguistics and Languages at Griffith University, Brisbane, Australia. He holds a B.A., A.Ed. from the University of Queensland and M.A. and Ph.D. in Applied Linguistics from the University of Essex, England. His principal areas of interest are in language policy-making, language education planning (including curriculum design and methodology) and second/foreign language testing (especially proficiency assessment). He has extensive experience of education (especially language education and language teacher education) in Australia and other countries, has published widely in books and journals published around the world, and is the co-author, with Elaine Wylie, of the International Second Language Proficiency Ratings, ISLPR (formerly known as the Australian Second Language Proficiency Ratings, ASLPR). He has had extensive consultancy involvements in many countries, was one of the development team for the IELTS test, and was Chief Examiner (Australia) from 1989 to 1998. In addition to his role in his Centre in Australia, he has been an Andrew Mellon Foundation Fellow and, since 1995, an Adjunct Fellow at the National Foreign Language Center, Washington DC, where he has been able to puruse his interests in national language policies.

References

Alderson, J. Charles and Arthur Hughes (eds.). (1981). *Issues in Language Testing, ELT Documents No. 111*. London: The British Council.

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

Bachman, L. F. and A. Palmer. (1996). *Language Testing in Practice*. London: Oxford University Press.

Clark, John L.D. (ed.). (1978). *Direct Testing of Speaking Proficiency: Theory and Application*. Princeton, N.J.: Educational Testing Service.

Coates, Sharon *et al.* (1995). *National Reporting System.* Canberra/Brisbane: Commonwealth of Australia and the Australian National Training Authority.

Education Committee, Council for Cultural Cooperation, Council of Europe. (1996). *Common European Framework of Reference for Language Learning and Teaching, Draft 1 of a Framework Proposal*. Strasbourg: Council of Europe.

Hyltenstam, K. and M. Pienemann. (1985). M*odelling and Assessing Second Language Acquisition*. Clevedon: Multilingual Matters.

Ingram, D. E. (1985). "Assessing proficiency: An overview on some aspects of testing". In Hyltenstam and Pienemann 1985: 215 - 276.

Ingram, D. E. (1985). *Report on the Formal Trialling of the Australian Second Language Proficiency Ratings (ASLPR)*. Canberra: Australian Government Publishing Service/Department of Immigration and Ethnic Affairs, 1984.

Ingram, D. E. and Elaine Wylie. (1991). "Developing proficiency scales for communicative assessment". In *Language and Language Education: Working Papers of the National Languages Institute of Australia*, Vol. 1, No. 1: 31 - 60. [ERIC FL019252/ED342209]

Ingram, D. E. and Elaine Wylie. (1979/1995). *The Australian Second Language Proficiency Ratings*. Brisbane: Centre for Applied Linguistics and Languages, Griffith University.[2]

Kellett, Marianne R. and Joy J. Cumming. (1995). "The Influence of English language proficiency on the success of non-English speaking background students in a TAFE vocational course." In *The Australian and New Zealand Journal of Vocational Education Research*, Vol. 3, No. 1: 69 - 86.

Lee, Tony. (1993). "A Many-Faceted Rasch Analysis of ASLPR ratings". Nathan: Centre for Applied Linguistics and Languages, Griffith University. mimeograph.

[2] The *Australian Second Language Proficiency Ratings* (ASLPR) was first published in January 1979 and has since passed through numerous editions including one published in 1984 by the Australian Government Publishing Service and, the most recent, published in 1995 by the Language Testing and Curriculum Centre in the Centre for Applied Linguistics and Languages at Griffith University, Brisbane, Australia, 4111. It also exists in numerous versions for different languages, for self-assessment, and for the assessment of proficiency in specified areas.

Messick, S. (1994). "The interplay of evidence and consequences in the validation of performance assessments". In *Educational Researcher*, Vol. 23, No. 2, March: 13 - 23.

Onions, C. T. (ed.). (1967). *The Shorter Oxford English Dictionary on Historical Principles*. Oxford: Clarendon Press.

Sollenberger, Howard E. (1978). "Development and current use of the FSI Oral Interview Test". In Clark 1978: 3 - 12.

Vollmer, Helmut J. (1981). "Why are we interested in 'General Language Proficiency'?". In Alderson and Hughes 1981: 152 - 175.

Weston, Kaye. (1995). "Language studies support for NESB students in AIS and Humanities, Report Phase 2, Semester 2, 1995." Nathan, Queensland: Centre for Applied Linguistics and Languages, Griffith University. Mimeograph.

Appendix One

The Australian Second Language Proficiency Ratings (ASLPR)

The Australian Second Language Proficiency Ratings (ASLPR) were initially developed by Elaine Wylie and D. E. Ingram in 1978 and first published in January 1979. The basic scale is designed to measure general proficiency or practical language skills in real-life language contexts in second or foreign language learners. The scale consists of 12 levels from zero to native-like, numbered from zero to 5 as shown below. The scale is presented in three columns: the first column provides a "General Description of Language Behaviour" and is almost identical across all versions of the scale, the second provides "Examples of Language Behaviour" and is specific to the particular version of the scale, and the third is a "Comment" column that explains, gives definitions and draws attention to critical features of the descriptor or level.

The outcome of using the ASLPR for the assessment of a second or foreign language learners' proficiency is a profile showing the rating for each macroskill separately, e.g., S:3, L:3+, R:2+, W:2. The levels in each of Speaking, Listening, Reading and Writing are identified with a number and a short descriptive title as follows:

0	Zero Proficiency	e.g., S:0, L:0, R:0, W:0
0+	Formulaic Proficiency	
1-	Minimum 'Creative' Proficiency	
1	Basic Transactional Proficiency	
1+	Transactional Proficiency	e.g., S:1+, L:1+, R:1+, W:1+
2	Basic Social Proficiency	
2+	(unnamed)	
3	Basic 'Vocational' Proficiency	
3+	(unnamed)	
4	'Vocational' Proficiency	e.g., S:4, L:4, R:4, W:4
4+	(unnamed)	
5	Native-like Proficiency	

Ingram and Wylie have worked on the ASLPR virtually continuously since 1978. It has been formally trialled in a number of different contexts and has been applied and re-developed in a number of different versions listed below. It is now the most widely used instrument for the specification of proficiency levels in Australia, is used in many places around the world, and has significantly influenced proficiency scale development elsewhere (e.g., the *ACTFL Guidelines*).

The ASLPR currently exists in the following versions:

- *The Australian Second Language Proficiency Ratings - Master General Proficiency Version (English Examples).* Brisbane: Centre for Applied Linguistics and Languages, Griffith University, 1995. ISBN 0 86857 814 2. Co-author Elaine Wylie

- *The Australian Second Language Proficiency Ratings - General Proficiency Version for English.* Brisbane: Centre for Applied Linguistics and Languages, Griffith University, 1979/1985/1995. ISBN 0 86857 815 0. Co-author Elaine Wylie

- *The Australian Second Language Proficiency Ratings - Version for Teachers of Indonesian.* Brisbane: Centre for Applied Linguistics and Languages, Griffith University, 1996. ISBN 0 86857 819 3. Co-authors Elaine Wylie and Geoff Woollams

- *The Australian Second Language Proficiency Ratings - General Proficiency Version for Indonesian.* Brisbane: Centre for Applied Linguistics and Languages, Griffith University, 1995. ISBN 0 86857 816 9. Co-authors Elaine Wylie and Geoff Woollams

- *The Australian Second Language Proficiency Ratings - Version for Second Language Teachers.* Brisbane: Centre for Applied Linguistics and Languages, Griffith University, 1995. ISBN 0 86857 817 7. Co-author Elaine Wylie

- *The Australian Second Language Proficiency Ratings - English for Business and Commerce Version.* Brisbane: Centre for Applied Linguistics and Languages, Griffith University, 1995. Co-authors Elaine Wylie and Hilda Maclean

- *The Australian Second Language Proficiency Ratings - English for Engineering Purposes Version.* Brisbane: Centre for Applied Linguistics and Languages, Griffith University, 1995. Co-authors Elaine Wylie and Laura Commins

- *The Australian Second Language Proficiency Ratings - English for Academic Purposes Version.* Brisbane: Centre for Applied Linguistics and Languages, Griffith University, 1995. Co-authors Elaine Wylie and Catherine Hudson

- *The Australian Second Language Proficiency Ratings - Version for Japanese.* Brisbane: Centre for Applied Linguistics and Languages, Griffith University, 1994. Co-authors Elaine Wylie and Peter Grainger

- *The Australian Second Language Proficiency Ratings - Version for French.* mimeograph. 1981. Co-authors Elaine Wylie and Edwige Coulin

- *The Australian Second Language Proficiency Ratings - Version for Italian.* mimeograph. 1981. Co-authors Elaine Wylie and Carlo Zincone

- Various self-assessment versions ranging from very short, simplified versions administered by telephone to computer-based versions, and versions used with language teachers.

Other versions are currently under development, e.g., a version for Korean and for the assessment of the proficiency of teachers of Korean.