Language Scales

D. E. Ingram

1. Introduction

This paper identifies some key issues relevant to the nature of scales and their relationship to 'real life performance' and to other assessment-related instruments. Rather than expound on current test theory, it seemed more appropriate for the present writer to discuss scales from the point of view of an applied linguist with an interest in language testing, with a great deal of day-to-day involvement in research into proficiency scales of different types, and with significant involvement in the academic maintenance and management of two large-scale tests (viz., IELTS and ACCESS).

2. All Assessment Instruments Are Compromises

Perhaps the most fundamental point to make is that all assessment-related instruments, whether scales, static competency specifications, observations, or tests 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 clients, 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. An essay-writing item in some tests of writing may be quite inappropriate and be said not to match real life but in an Academic Purposes test it may do so.

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 occur in real life language use.

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; Wylie and Ingram 1995] seeks to describe language proficiency as it develops from zero to native-like and to be useable 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 the 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. In addition, as verbal, written descriptions rather than videotapes of language use, they 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 make the match between the verbal descriptions and 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 of the scale an image of the underlying behaviour that is realised in each real life instance of language use. In brief, scales such as the ASLPR are not 'real life' but they seek to suggest real life language behaviour and are validated against real life language behaviour or performance.

In discussing 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. 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 in which recorded prompts are used to elicit the learners' language for recording on tape and subsequent assessment. 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 and to allow test method to intrude less than in other 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). The interviewer is instructed to use activities that serve several purposes including to so structure the interview as to lead learners on to demonstrate their maximum language performance and to provide a variety of tasks within the space of a 20 to 40 minute interview while at the same time ensuring that, for the learner, the interview is as relaxed as is appropriate, that it flows naturally, and that it uses real life language tasks. Nevertheless, the interview is a compromise with real life though one whose method is designed to intrude less on the reality of the learner's 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 descriptions match what they know to be their 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.

3. Cooperation Or Conflict: Scales And Tests

In introducing this paper, the term 'assessment-related instruments' was used. The reason for this was simply that scales are not 'tests' in the traditional sense of the word and are not necessarily used for assessment purposes. Nevertheless, many scales such as the Australian Second Language Proficiency Ratings (ASLPR) [Ingram and Wylie 1979/1985; Wylie and Ingram 1995] are available for use in assessment and are also used for other purposes such as providing an overarching framework within which to plan language programmes and specify entry levels and exit goals. Even in the assessment context, scales and the procedures used to elicit language behaviour for assessment against them are different.

One corollary of these observations is that it would be simplistic to say that scales and tests are in competition: they serve different purposes and, even when used within the context of assessment, are appropriate to different purposes in different contexts of use.

In addition, however, scales and other assessment instruments are complementary and may be used cooperatively in that scales can be used to explicate and assist in the interpretation of test results while tests of a variety of different types may be designed specifically to assign learners to levels on a scale. The IELTS and ACCESS tests illustrate the complementary and cooperative use of scales in major tests since, in these batteries, proficiency in speaking and writing are assessed 'directly', with learners' language behaviour in these macroskills being observed and matched against scale descriptors while their listening and reading abilities are assessed with tests that take a variety of different forms generally using so-called objective item-types, including (though not only) multiple choice, and the scores are then related to levels on simple proficiency scales. Similarly, though one tends to think of the ASLPR being used for proficiency assessment in the context of eliciting language in an interview, it can also be used where proficiency assessment is conducted by observations *in situ* in real life. It is also used to explicate or interpret other test results, ie., to express the results of more traditional, standardised tests in proficiency terms.

The problem of how results on more traditional tests are related to descriptive scales has not, in practice, been resolved. In principle, it

should be possible to provide test specifications that enable a test to be graduated according to the progression on a descriptive scale so that learners would proceed successfully in a test up to their proficiency level. In practice, the complexity of language and the complexity of the factors that lead to a learner's successfully carrying out any test task are such that no such simple relationship between item performance and proficiency scale levels usually occurs even though one might expect from item response theory that it should be possible to tie items quite closely to proficiency levels. In practice, in IELTS for example, the link between performance on listening and reading and the bandscale is made in an almost entirely norm-referenced manner by distributing test results over nine levels, comparing performance with performance of all previous cohorts on anchor items and previous versions of the test, and then assigning bandscores on the nine level IELTS bandscales. The link between item performance and bandscale scores is further attenuated by providing an 'overall' bandscore which is more or less the mean of the four macroskill-specific bandscores; contrary to the preferred advice of the test developers, it is this overall score that most institutions use. In brief, though, it should be possible, in principle, to link test specifications closely to scale levels, in practice it is not and the use of scale descriptors to explicate test results entails a massive compromise in the area of the validity of the interpretation of test scores but a compromise intended to assist naive end-users to interpret the otherwise fairly opaque numerical test results. The fact of the matter is that most end-users are ultimately interested in the learners' practical language skills (ie., what tasks they can carry out in what sort of language) whereas tests characteristically produce numerical results of some sort that are more related to comparisons between candidates than to statements about language behaviour; the use of scales to interpret test results is intended to assist end-users to make this interpretation.

One might argue that this somewhat stochastic business of interpreting test scores in terms of performance statements on proficiency scales might be less insecure if test performance and scale-based ratings were also correlated with real life performance such as, for example, candidates' ability to carry out language-related tasks in an academic course, but studies of predictive validity are notoriously difficult to structure, their results are almost invariably subject to criticism because of the multitude of

uncontrollable or barely controllable variables involved in real life language-based performance, and such correlational studies are rarely attempted and even more rarely convincing. Thus they do not provide a realistic means by which to give greater security to the scale-based interpretation of test results no matter how interesting they might be from a language testing point of view and no matter how useful end-users might assert such studies would be for their purposes.

4. 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, ie., 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, ie., 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 the other skills the particular workplace requires. The recently released National Reporting

System compiled by Sharon Coates and others [1995] seems to take this form in application to language and literacy competencies.

In earlier papers, Elaine Wylie and the present writer identified several different types of proficiency scales according to how they are constructed, what they attempt to measure, and what criteria they contain [Ingram and Wylie 1991]. Without pursuing all of the issues in any detail, the contrasts identified included:

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 or may select only certain specific tasks which are graduated in some way along a scale, as is exemplified in some vocational competency specifications. This issue is further discussed subsequently.

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 as 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 one or more of the macroskills separately or, like the oral interaction scale for the ACCESS test, they may seek to describe the combined behaviour of two or more macroskills, or, as in the overall IELTS bandscale, they describe general language behaviour in a way that is supposed 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 language development and claim to provide a global picture of language behaviour. These scales accept that some of the parameters of change may develop rapidly within the total skein of language behaviour and that it is the total picture of language behaviour that is more practically significant than the fact that certain parameters are more or less developed than the characteristic development pattern suggests at the particular level in question.

Analytic vs Holistic: Some scales may seek to provide quite detailed statements about, for instance, specific aspects of grammatical development at particular levels on the scale. Other scales 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.

Elaine Wylie has a useful typology of scales as used in language assessment and this is shown in Table 1.

All of these contrasts warrant extended discussion but here reference will be made to just two that are particularly relevant, viz., what above has been called task-only vs total or underlying behaviour, as seen in the contrast between the specification of competencies and the measurement of general proficiency in, for instance, the ASLPR. Reference will also be made to the somewhat related issue of proficiency vs. course achievement.

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 and it is for this reason that instruments such as the ASLPR insist that if one is seeking to measure that underlying ability it is essential to use more than one task. In a task-centred approach, as found in the specification of some 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, and 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]

As already noted, the distinction between competency specifications and their use in performance assessment, on the one hand, and general proficiency and its assessment using proficiency scales, on the other, is somewhat analogous to the distinction between course achievement and the assessment of general proficiency. In principle, the content of a course could be ordered in some sort of scale or it might be identified and assessed without any particular ordering. In one sense, course achievement and its assessment relates to assessment of prior learning and performance on a particular course,

whereas general proficiency and its assessment are not specified against a particular course, its content and its mastery but more independently so that, again, it is the underlying ability that is measured rather than knowledge of and performance on specific course content. In this sense, one can contrast course achievement which is backward-looking to the course with general proficiency assessment which is forward-looking to subsequent performance beyond the test and beyond how the proficiency has been acquired. Course achievement is given in terms of quite specific course components whereas proficiency assessment, such as with the ASLPR, is presented on a scale representing a continuum of general language development and representing an underlying and growing ability, irrespective of how that ability has been acquired.

5. 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 abilities 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 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.

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 that, amongst other things, provide a common language by which to talk about the otherwise nebulous concept of language proficiency.

In developing the ASLPR (an activity which has proceeded now for some 17 years since 1978), the following actions were taken:

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 these tasks.

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.

The initial descriptors were then tested out, elaborated and refined in interviews with learners throughout the proficiency span in which the features of their language were deliberately elicited 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 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.

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.

The scale was formally trialled using adult and adolescent learners, especially of English but also of other languages. 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.

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 and the Language Testing and Curriculum Centre 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, first, the levels in the scale actually do represent a progression from zero to 5 along a common dimension and, second, 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. The ASLPR scale and its sub-scales seem able to uncover an ESL proficiency developmental path of learners from diverse L1 backgrounds and age groups with data covering two years...

6. There is little noise in the system. (In practice what this meant was that, of the more than 300 candidates, 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.

6. Validity And Reliability

Many of the issues already discussed, especially in relation to the development of scales, touch on issues of validity and reliability and, in this brief paper, will not be discussed at length again 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 are, at best, difficult to assess or reject but the validity of scales such as the ASLPR are 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 what is known from psycholinguistics about how language develops. In summary, however, it has to be said that the validity of a scale can be assessed only against what is know about what it seeks to describe, its purpose, the context in which it is to be used, and the context within which the findings are to be used. Reference has also been made already to the usefulness of predictive validity studies in assessing the extent to which scales and assessments using them reflect real life language performance, but it was also noted that it is rare that such studies are adequately structured or convincing. 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.

As already noted, both in the development process and in the validation studies, a great deal of the work on the ASLPR was directed at identifying the key features of language development both throughout the proficiency span and at the various defined levels within it and ensuring that the descriptions provided were consistent and coherent both from level to level and within each level. Without wishing to be unduly critical of an instrument that has only recently been developed and released and hence has not been subjected to prolonged research, one might note that such consistency and coherence seems to be lacking from the National Reporting System [Coates et al 1995], perhaps because, as static specifications of competencies, it seems to lack any underlying notion of how language and performance develop. This may be appropriate if its aim is just statistically to identify competencies but problems arise if it takes on developmental overtones or if it is assumed that elements in each level are somehow developmentally related, as would inevitably occur if it is used to make ability statements about second language learners or is used to sequence curricula. The problem is illustrated in Level 3 (Reporting Information). Here, the 'oral communication' indicator of competence states in 3.7:

> "Takes part in short inter-personal exchanges, demonstrating some awareness of register and interactional strategies, for the purpose of establishing, maintaining and developing relationships; exploring issues; or problem-solving"

while, under 'Vocabulary and Grammar', the descriptor states:

> "Uses and comprehends simple grammatical forms and vocabulary to give instructions, give explanations, ask questions, and express viewpoints...

In further contrast, under 'Public Communication—Speaking and Listening' in Level 3, it is stated:

> "Restates the main idea of a text and evidence offered in support of this view, after viewing or reading persuasive text(s), eg., TV advertisements, public notices, political advertisements."

> Discusses the content after reading an article in the daily newspaper/viewing TV program...

Clearly, if applied to second language learners, there are considerable differences in the linguistic and psycholinguistic expectations of these different elements of performance within Level 3 and the most one can say is that there is an assumption that competencies can be identified and somehow grouped without account being taken of how they are developed or their compatibility in terms of assumed proficiency levels or developmental stages. If this is intentional and justifiable in terms of the purpose and context of use of the NRS, then the NRS's validity may not be in question but, if it is to be used, in practice, to state where second language learners are in their continuum of development, one would have to question whether it can meet minimum validity requirements for such an instrument. What is not yet clear, because the research has not been done, is whether it is possible to overlay a proficiency scale such as the ASLPR over the NRS, in order to add a proficiency development dimension to its levels but, at first glance, this does not seem likely.

Reliability of scales seems to hinge on the extent to which different users can interpret them in the same way. Again 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, 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' training sessions 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.

7. Conclusion

This paper has looked at just a few of the issues to be considered in relation to scales. Both scales and other assessment instruments attempt to make statements about aspects of a person's ability to use language or the features that compose language. All assessment-related instruments are compromises that serve different purposes in different contexts of use and their usefulness, their appropriateness and their validity and reliability can be assessed only if one takes account of those purposes and contexts. Scales and other assessment instruments are, therefore, not necessarily in competition or in conflict but are complementary and may be mutually supportive. Proficiency scales attempt to provide a common yardstick allowing statements to be made over a specified range, commonly zero to native-like. As such, they are also used to interpret results on other instruments even though the manner in which such interpretation occurs often gives greater acknowledgment to end-users' practical needs than it does to such 'theoretical' issues as validity.

8. References

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

Hyltensam, K. and M. Pienemann (1985) *Modelling 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. and Elaine Wylie (1991) 'Developing proficiency scales for communicative assessment' In *Language and Language Education: Working Papers of the National Languages Institute of Australia,* 1,1: 31–60 [ERIC FL019252/ED342209]

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' Mimeograph

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

Messick, S. (1994) 'The interplay of evidence and consequences in the validation of performance assessments' In *Educational Researcher,* 23,2, March: 13–23

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

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

The author wishes to thank a number of colleagues who have contributed ideas to this paper, in particular Elaine Wylie, Laura Cummins, and Cath Hudson.

The author was a member of the team that developed IELTS in 1987-89 and has been Chief Examiner (Australia) since 1989. The ACCESS test is the test used by the Australian government with applicants for immigration to Australia. It is administered overseas by IDP Education Australia and the academic management is carried out by the Centre for Applied Linguistics and Languages in Griffith University, Brisbane, Australia. The author is the Director of the Centre and has a major editing role in the development of new versions of ACCESS. He is also co-author (with Elaine Wylie) of the Australian Second Language Proficiency Ratings (ASLPR) and has been continuously involved in research on the ASLPR and proficiency assessment since 1978.