

An Overview of the ISLPR®

(International Second Language Proficiency Ratings)

Elaine Wylie

2010

Further information about any of the areas outlined can be obtained from the author at ISLPR LANGUAGE SERVICES PTY LTD.

Phone/fax +61 (0)7 3423 2505

Email e.wylie@islpr.org

The trademark ISLPR® is owned by Co-Directors of ISLPR LANGUAGE SERVICE, David Ingram and Elaine Wylie.

CONTENTS

	page
1 The nature of the ISLPR	1
2 Its uses	2
2.1 Assessing the proficiency of individual learners	2
"*****Cuugukpi "d{ "vgukpi	" 2
"*****Cuugukpi "d{ "o gcpu'qyj gt'yj cp'vgukpi	***** 4
2.2 Research and policy-making	4
2.3 Providing a framework for language curriculum development	5
3 A very brief history of the development of the scales	6
4 Validity	6
Concurrent validity	6
Predictive validity	7
Construct validity	7
5 Reliability	8
5.1 Inter-rater reliability	8
5.2 Intra-rater reliability	9
5.3 Test-retest reliability	10
References	10

1 The nature of the ISLPR¹

The ISLPR[®] is a scale that describes the development of second or foreign language proficiency in adolescent and adult learners. More precisely, it is a set of four subscales for the macroskills of speaking, listening, reading and writing. These subscales trace the development of the target language from 0 (no ability to communicate in the target language) to 5 (indistinguishable from a native speaker of the same sociocultural background). There are intermediate ‘plus’ or ‘minus’ levels, giving a total of 12 levels in each subscale.² The names of the levels and a very simple descriptive statement are given below.

0	ZERO PROFICIENCY	Unable to communicate in the language.
0+	FORMULAIC PROFICIENCY	Able to perform in a very limited capacity within the most immediate, predictable areas of need, using essentially formulaic language.
1-	MINIMUM ‘CREATIVE’ PROFICIENCY	Able to satisfy immediate, predictable needs, using predominantly formulaic language.
1	BASIC TRANSACTIONAL PROFICIENCY	Able to satisfy basic everyday transactional needs.
1+	TRANSACTIONAL PROFICIENCY	Able to satisfy everyday transactional needs and limited social needs.
2	BASIC SOCIAL PROFICIENCY	Able to satisfy basic social needs, and routine needs pertinent to everyday commerce and to linguistically undemanding ‘vocational’ fields.
2+	SOCIAL PROFICIENCY	
3	BASIC ‘VOCATIONAL’ PROFICIENCY	Able to perform effectively in most informal and formal situations pertinent to social and community life and everyday commerce and recreation, and in situations which are not linguistically demanding in own ‘vocational’ fields.
3+	BASIC ‘VOCATIONAL’ PROFICIENCY PLUS	
4	‘VOCATIONAL’ PROFICIENCY	Able to perform very effectively in almost all situations pertinent to social and community life and everyday commerce and recreation, and generally in almost all situations pertinent to own ‘vocational’ fields.
4+	ADVANCED ‘VOCATIONAL’ PROFICIENCY	
5	NATIVE-LIKE PROFICIENCY	Proficiency equivalent to that of a native speaker of the same sociocultural variety.

The full scale, including introductory notes and a glossary, is a 50-page document. Most of the levels are described in detail (i.e. a page of rich description for each level in each macroskill). The description includes a statement of the kinds of tasks that people at that level can perform (with the contexts they can perform them in) and the kinds of language forms they use when performing those tasks (with detail about accuracy, range, fluency, appropriateness, etc.). The descriptions assume real-life, communicative language use.

¹ The name was changed by developers David Ingram and Elaine Wylie from ASLPR (Australian Second Language Proficiency Ratings) in 1997 to reflect the increasing international use of the scale.

² There are some exceptions to this in versions for professional purposes for which the project advisory committee considered that to include descriptions of the lowest levels might be interpreted as legitimising the employment of learners with very low proficiency.

There are a number of versions of the scale. They can be divided into two groups: *general proficiency* versions and *specified purpose* versions. The *general proficiency* versions facilitate *learner-focused* perspectives; that is, when the focus is on individual learners' own domain-related strengths in the language, which depend on the individual's personal experiences of the language in real-life situations and/or formal studies. There are general proficiency versions for different target languages, and a generic 'master' version, which can be used for any language.

The *specified purpose* versions facilitate *role-focused* perspectives; that is, when the focus is on a particular domain or domains of language use relevant to a particular role. This is appropriate when some party – usually a 'consumer' or end-user of test results such as a professional registration board, an employer, a court of law, or a dean of studies in a tertiary institution – specifies that they are interested in levels of language ability in a particular set of domains. The *specified purpose* versions developed so far have been for academic purposes and a range of professional purposes.³

The page format is the same for all versions. There are three columns. The first column describes broad semantic, linguistic and sociocultural phenomena; it is almost the same for all versions of the scale. The middle column is language specific and, in the case of the specified purpose versions, domain specific; it features examples of tasks that are appropriate to the target culture and/or the broad context of language use (i.e. whether in a second or foreign language situation) and of language forms. The third column, which has a high level of commonality across versions, highlights the main changes in behaviour observable at the particular level and provides a commentary on phenomena described at the level.

2 Its uses

The ISLPR is used for three broad purposes: (i) assessing the proficiency of individual learners; (ii) research and policy-making and (iii) providing a framework for language curriculum development.

2.1 Assessing the proficiency of individual learners

The orthodox process of assessing (or rating) a learner is a direct, holistic one of matching observed behaviour against the descriptions of the scale. People can be assessed by *test*⁴ or *non-test* means.

Assessing by testing

Orthodox *learner-focused* approaches to testing are fully adaptive to the candidate in terms of his/her communicative needs and interests and his/her proficiency level (as hypothesised and continually re-hypothesised by the tester as the test progresses). Listening and reading as well as speaking are tested in a face-to-face interview; the dynamic nature of this test method is in accord with interactive theories of listening and reading. The interview is usually one-to-one but in very high-stakes situations a second person may be involved. This person acts as an observer who, if the candidate is demonstrating listening skills of at least Level 1+, can also participate in appropriate

³ Copies of some of the versions can be ordered from ISLPR LANGUAGE SERVICES, info@islpr.org

⁴ Application forms for tests can be found on the ISLPR LANGUAGE SERVICES webpage www.islpr.org
For information about training to become a tester or about quality assurance processes and licensing, please contact the author.

parts of the interview.⁵ Writing tasks may be done in a room with other candidates (who will very likely have different tasks) with an appropriate level of supervision which allows, *inter alia*, for the candidates to seek help if they do not understand the nature of any task.⁶

Role-focused tests range from those that are relatively standardised (in the general rather than the statistical sense), with specifications developed in consultation with the end-users of test results, to those that are adaptive but concentrate on the relevant domain(s).

This can be seen in terms of a ‘strong-weak’ continuum. Typical of the stronger forms of the test are those that are used for professional registration. In tests that are administered to overseas-trained professionals seeking registration in Queensland, the tester must not only access domains that relate to the roles of a teacher as *practitioner* and *professional in society*, but must also use certain task types. For example, reading tasks for overseas-trained teachers must include a simulation in which the candidate reads aloud a text (e.g. a notice from the school principal or a segment of a ‘big book’) to the tester, as if to a class. One-off strongly role-focused tests have been conducted by the present writer for a wide range of purposes. For example, a defendant in a fraud case was tested in reading and writing; at issue was whether this person had understood specifications for a building project, and the test focused on technical and legal language. Another test was administered to welders operating in confined spaces at the request of their employer, a ship-building company; it focused on the oral and written language of industrial safety.

Typical of weaker forms would be when the candidate has given as the reason for doing the test that s/he is seeking entry to academic studies that have no pre-requisite discipline studies. The tester chooses texts and tasks that reflect the candidate’s interest in the discipline but do not require any specialised language. For example, for a candidate who wishes to enter an MBA program, one of the reading tasks might be a story about management practices from the general section of a daily newspaper. The tester must then decide whether to narrow the focus (for example, with an editorial from a weekly business newspaper or magazine) or to choose a text that reflects the candidate’s background in (say) engineering or political science. The latter choice would represent an interface with a *learner focused* approach.

There is also a continuum in terms of the degree to which technical language needs to be accessed. If consumers consider technical language necessary and the testers are not specialists in the focal domain(s), there is a case for specialist input.

⁵ The participation of the second person relates to the validity of the testing process; according to Level 2 in the listening subscale, “the presence of other participants in the conversation does not *per se* cause problems”. Rating accuracy is not compromised if it is not practical for a second person to be present; all interviews in high-stakes situations should be video- or audio-recorded for a second judge to rates the candidate’s performance in speaking, listening and reading on the basis of the recorded data. If there is any difference between ratings, this should be resolved through consultation or the involvement of a third, independent expert.

⁶ In high-stakes situations the script is rated by a second person and any difference between ratings is resolved as per Footnote 5.

Assessing by means other than testing

Learners can assess their own proficiency by reflecting on their own recent real-life use of the focal language and matching these reflections against one of the simplified versions of the ISLPR subscales designed for self-assessment (e.g. Wylie, 1997). One very simple version was developed for a major telephone survey of community language resources in Australia; another is programmed for computer administration. Self-assessment scales have been translated into a number of learners' first languages to make them more accessible. One version (for learners of Japanese who may have had little experience in communicative use of the language) is complemented by example tasks with 'answer keys' to help learners to interpret their performance.

Assessment can be made by external raters in naturalistic situations. Judgements made on the basis of non-intrusive observations are particularly relevant to high-stakes *role-focused* situations where key abilities would be difficult to elicit in a test. The present writer has assessed candidates on the basis of her observations of real-life interactions between non-English-speaking background (NESB) and native-speaking adults in a wide variety of settings, including legal chambers⁷, a prison, a call centre, a school, and a university campus.

ISLPR reporting is in the form of a profile, with speaking, listening, reading and writing indicated separately (e.g. S:2+ L:2+ R:3 W:2). It is not appropriate to give a candidate an 'overall' (aggregated) rating. Decisions about key proficiency levels (e.g. minimum levels for entry to academic courses) are generally made by the end-users in consultation with experts who provide samples of language behaviour to aid in the interpretation of the subscales.

2.2 Research and policy-making

When the ISLPR is used for purposes related to research or policy-making, the ratings of individual learners are often aggregated. For example, Gariano (1997) conducted a large-scale study of the role of English in the settlement of immigrants, using ASLPR ratings of 32,735 clients on entry to and exit from English language provisions in the Australian Adult Migrant English Program (AMEP). Lee *et al* (1998) conducted a complementary study of 8,094 AMEP clients in Queensland.

Aggregated ISLPR data have been used in evaluations of language programs in Indonesia (Phillips *et al*, 1985), Australia (e.g. Ingram *et al*, 1996) and East Timor (Australian Foundation for the Peoples of Asia and the Pacific, 2000).

A number of community-based surveys have used the scale. These include studies of the English needs of immigrants to Australia (e.g. Smith and Baldauf, 1982; Manton *et al*, 1983; Coppell *et al*, 1984) and a survey of skills in languages other than English in Australia (Ingram and Lee, 1993).

The ISLPR has been used in audits of language resources and needs analyses in the vehicle manufacturing industry in Australia (Sefton and O'Hara, 1992) and an Australian-owned mining venture in Indonesia (Singh *et al*, 1999).

⁷ ISLPR assessments for forensic purposes are normally done by means of a test (e.g. Gibbons, 1990; Jensen, 1995). In the legal case referred to here, the naturalistic setting was considered more conducive to revealing the true proficiency of a defendant who might have been tempted to 'malinger' linguistically.

More recently (2009-10) it was used in AusAID-funded audits of the English proficiency of educators in the Republic of Kiribati. Approximately 1500 administrators and curriculum officers in the Department of Education, lecturers at the Kiribati Teachers College and the Kiribati Institute of Technology, teachers and trainee teachers were tested and are being retested after intervention programs.

The scale has been used as the yardstick in a number of reports commissioned by Australian state and federal governments and in policy statements on language and language education (e.g. Ingram and John, 1990; Australian Language and Literacy Council, 1996). It was one of the inputs to the UN's ICAO (International Civil Aviation Organization) Language Proficiency Requirements for non-English-speaking background pilots and air-traffic controllers (Matthews, 2004) and was used in the validation of Griffith University's ALITE test, which was developed in accordance with ICAO's requirements.

In 2009, Australian state and territory transport regulatory authorities recognised a need to establish minimum standards of English for NESB taxi drivers and chose the ISLPR as the instrument to be used. To identify tasks that drivers perform, researchers from ISLPR Language Services listened to recordings of driver-passenger interactions; interviewed drivers, managers and trainers from taxi companies and departmental officers; observed driver-operator interactions from radio control rooms; and perused official documents and media reports. The levels recommended, *viz* S:2+, L: 3, R:2, W:1+, were accepted (Ingram, 2010; Ingram and Wylie, 2010).

Scores on other proficiency tests have been mapped on to ISLPR levels (e.g. UNSWIL, n.d.; Lee, 1992; NLLIA, 1996; Mousavi, 2007) as have competency-based reporting systems (Ingram and Wylie, 1996).

2.3 Providing a framework for language curriculum development

The ISLPR levels provide a framework for a language program, guiding administrators and curriculum developers and teachers in overall program and course planning for English in second and foreign language learning situations (e.g. Ingram, 1979; Wylie, 2000) and other languages in schools and universities (Ingram and Wylie, 1991; Willcock, 2001). A number of language teacher education courses are planned around ISLPR levels as outcomes (e.g. Tasmanian Educational Consortium, 1997; Griffith University, 2000). Cross-language comparisons in terms of progress are valid because of the fundamental sameness of (say) Level 3 in speaking in all languages.

The scale also provides guidelines for curriculum developers and teachers in terms of the development and selection of teaching/learning tasks (Department of Immigration and Ethnic Affairs, 1979; Wylie, 2000) and materials (Beaumont, 1982/83; Beaverson and Carstensen, 1982). The level above the learner's actual level in each macroskill is set as a goal⁸. The teaching/learning program focuses on the types of tasks that learners at the higher (goal) level can perform, and activities and materials are chosen so as to optimise the learner's progress towards that higher level. The amount of 'scaffolding', or support provided to enable learners to perform the higher-level tasks (Wood, Bruner and Ross, 1976) can be varied to take account of individual differences in proficiency or preferred levels of challenge. This developmental approach is very similar to that proposed by Vygotsky (1978).

⁸ In long-term or articulated programs this next level is likely to be an interim or 'waystage' goal.

3 A very brief history of the development of the scales

The Joint Commonwealth-States Committee on Professional Aspects of the (then) Adult Migrant Education Program initiated the development of the (then) ASLPR in 1978. They needed an assessment instrument for policy and curriculum development and were committed to having an instrument that would reflect prevailing and emerging theories of adult learning, language, and second language learning (JCSC, 1979). The (then) academic adviser to the Committee, David Ingram, proposed the development of an Australian proficiency scale drawing on an existing US instrument, *viz* the Absolute Language Proficiency Ratings developed by the US Foreign Service Institute School of Language Studies (the FSI scales) (1968). This FSI instrument consisted of two subscales (for speaking and reading), each of which had five described levels. Ingram's proposal was accepted, and he was asked to lead a Queensland working party that would take the FSI scales as their starting point, but make necessary changes.⁹ These included two extra subscales (for listening and writing) and greater definition at the lower end of the scale to reflect the language abilities of the majority of NESB migrants at that time (two extra described levels between 0 and 1 and one extra described level between 1 and 2).

The first version of the ASLPR was released in 1979 for trialling by teachers in the field. Formal trialling for validity and reliability of the scales took place in 1980/1 (see below). The revised scale, incorporating changes resulting from the trials, was published in 1984 (Ingram and Wylie, 1984).

The 1995 version and subsequent versions of the ISLPR are the result of observations of thousands of learners by Ingram and Wylie and their colleagues, and continual reworking of the descriptors to make them reflect as closely as possible the broad developmental path of second/foreign language learners and to do so in a way that is helpful to users of the scale. Reference is made below to studies that have provided data on the validity and reliability of these more recent versions.

4 Validity

The initial trialling of the scale, funded by the (then) Department of Immigration and Ethnic Affairs, took place in Australia in 1980/81. To determine *concurrent validity*, Ingram and Wylie compared their ratings of 18 adult and adolescent migrant learners who represented the full proficiency span from *zero* to *native-like* with the learners' scores on the Comprehensive English Language Test (CELT) and on cloze and dictation tests. The following is a summary of Spearman correlation coefficients:

ASLPR macroskills and CELT (total of tests of structure, vocabulary and listening)

Speaking: 0.90 Listening: 0.88 Reading: 0.95 Writing: 0.96

ASLPR macroskills and cloze (total of three passages)

Speaking: 0.89 Listening: 0.85 Reading: 0.93 Writing: 0.94

ASLPR macroskills and dictation (total of three passages)

Speaking: 0.90 Listening: 0.88 Reading: 0.97 Writing: 0.94.

⁹ More recent US scales derived from the FSI scales, *viz* the Interagency Language Roundtable (ILR) Language Skill Level Descriptions (1985) and the American Council on the Teaching of Foreign Languages (ACTFL) Proficiency Guidelines (1986), have in turn drawn heavily on the ASLPR/ISLPR.

In all cases, the level of significance was at or beyond 0.001. The data suggest a high level of concurrent validity (Ingram, 1984).¹⁰

Soon after the initial trialling, a study funded by the (then) Australian Development Assistance Bureau attested to the *predictive validity* of the ASLPR. Phillips *et al* (1985) followed 76 Indonesian post-graduate students through to the end of the first year of their studies in Australian universities. Their results confirm that ASLPR Level 3+ in all macroskills is an ideal pre-requisite for students entering most post-graduate courses but that government-sponsored students in a very supportive study environment who were Level 3 on entry “were able ... to cope with the requirements of their courses.” (p.21)

Kellett and Cumming (1995) followed 35 NESB migrant students in a Technical and Further Education (TAFE) course, Certificate in Commercial and Office Fundamentals. They found that students needed at least ASLPR 2+ in all macroskills to succeed in this course, and those with at least Level 3 were much more likely to succeed. (p.83). They concluded that “language proficiency on entry was strongly related to academic success and could not be compensated for by other positive attributes of the students.” (p.71)

Most recently, Sefton and Wylie analysed the results of NESB international students who entered Griffith University *via* an ISLPR test in 1998, 1999 and 2000. No students who were accepted in accordance with Griffith policy (the levels required at the time were generally 3 in all macroskills for undergraduate courses and 3+ in all macroskills for postgraduate courses) received more than one grade of Fail or Pass Conceded in their first semester (Sefton and Wylie, 2002).

A number of studies have attested to the *construct validity* of the ASLPR/ISLPR scales. Lee (1992) used a many-faceted Rasch program (Linacre and Wright, 1990) to analyse the ratings in the ASLPR Testing Service database at Griffith University. From the establishment of the Testing Service in 1990, there had been 329 clients, the majority of whom were international students (others were NESB Australian residents). They came from over 10 different language backgrounds and ranged in age from 14 to 40. Most had taken the test because they were seeking entry to secondary or tertiary education in Australia; some needed to demonstrate their proficiency for professional registration. There were a total of eight tester/raters. Lee found that, at a significance level of 0.01, no ratings were misfitting, and at a level of 0.001, only one rating in one macroskill was misfitting. He concluded, *inter alia*, that:

- the ordinal nature of the ASLPR levels was established
- the nature of the four macroskills as subscales of the ASLPR was established
- the ASLPR and its subscales seemed able to uncover a proficiency development path of learners of English as a Second Language from diverse backgrounds and age groups.

¹⁰ When Ingram replicated the concurrent validity study in China, with students following a very traditional grammar-translation approach to learning English and lacking the opportunity to use the language for communicative purposes, the agreements were much weaker. For example, the Spearman correlations between ASLPR ratings and CELT scores of third-year students at the Guangzhou Institute of Foreign Languages ranged from 0.30 (ASLPR reading and CELT) to 0.58 (ASLPR Listening and CELT). These students' passive knowledge of the language (as tested by the CELT multiple-choice format) was high but, unlike the migrant learners in Australia, they could not mobilise this knowledge effectively to perform communication tasks.

Gariano (see 2.2 above) used a multiple-regression model to analyse records of 32,735 AMEP clients enrolled in English language programs around Australia in 1991. In his discussion of the data set he notes that he found ‘errors’ (i.e. initial ratings greater than subsequent ratings) in fewer than 2% of cases and concludes “This low error level suggests that the ASLPR scale adequately discriminates a person’s English language proficiency.” (Gariano, 1997:184).

One of the major threats to the construct validity of tests is “construct underrepresentation”, which in turn threatens the degree to which one can generalise beyond a candidate's behaviour on the test (Messick, 1994: 14-15). In performance tests that involve relatively extended tasks, Messick stresses the importance of achieving an appropriate balance between “breadth of content coverage and the depth of process understanding” (p15). Orthodox tests used to rate a learner on an ISLPR subscale, particularly at higher levels of proficiency, often involve tasks that are relatively extended. For example, at middle to upper levels a typical writing task would be a 250-word report, memorandum, open letter, or essay; a listening task might involve a self-contained segment of a lecture or a documentary program lasting five minutes. Practicalities constrain the number of such tasks that can be used, so that there is potential for breadth being sacrificed to depth.

To investigate whether this aspect of construct validity is compromised in the ISLPR testing system, Wylie has compared ratings made on the basis of performance in tests involving the relatively small number of tasks that constitute a formal test with those made on the basis of learners’ self-assessment of their performance in what has been in many cases a vast number and variety of real-life tasks. The learners rated themselves on simplified versions of the ISLPR subscales, translated where necessary into their first language. Studies have been conducted with English, Chinese and Indonesian as target languages. With learners of English, the level of agreement between test ratings and self-assessments for speaking ranged from 0.89 (when the learners were international students in intensive English courses) to 0.61 (when the learners were in a foreign language situation where they had had little or no opportunity to use the language) (Wylie, 2001). These compare very favourably with other studies where students have self-assessed on (simplified versions of) the same scale as tester-raters have used. For example, Clark and Swinton (1979), using the FSI scales, found a correlation of 0.48 with ratings for speaking, and Wilson (1996), using the ILR scales with a sample of learners which contained a high proportion of educated Swiss learners of French and German, found correlations “in the region of .70 throughout” (reported in Oscarson, 1997, p180).

Test-retest data (discussed below under Reliability) also suggest that the validity of ISLPR ratings does not suffer from construct underrepresentation.

5 Reliability

The 1980/81 trialling of the ASLPR addressed both inter- and intra-rater reliability. Test-retest reliability has subsequently been demonstrated in training programs and in the field.

5.1 Inter-rater reliability

In order to determine inter-rater reliability, Ingram and Wylie presented the language samples of 16 of the learners (using videotapes and photocopied scripts) in random order to 21 teachers in the AMEP in Sydney, Melbourne and Adelaide. The teachers’ training in the ASLPR, as self-reported on information sheets they filled out, ranged from “none” to

“extensive study of testing including ASLPR” (Ingram, 1984: 67-69). The following is a summary of the level of agreement between the teachers and the authors, using Spearman correlation coefficients:

Speaking

Correlations ranged from 0.91 to 0.99, with a mean of 0.96

Listening

Correlations ranged from 0.89 to 0.98, with a mean of 0.94

Reading

Correlations ranged from 0.91 to 0.99 with a mean of 0.96

Writing

Correlations ranged from 0.93 to 0.99, with a mean of 0.96.

In all cases, the level of significance was at or beyond 0.001. The data suggest a high level of inter-rater reliability. There was no indication that learners at the levels that are not fully described, viz 2+, 3+ and 4+, were rated more or less reliably than those at the other levels (Ingram, 1984).

Field trialling has also demonstrated a high level of inter-rater reliability. In 2005 Wylie analysed double ratings of the most recent 250 tests in the ISLPR Testing Service at Griffith University. The initial tester’s ratings for each candidate were compared with those of another tester who judged the candidate’s performance in speaking, listening and reading on the basis of audiorecordings and writing on the basis of the original, non-annotated scripts. Pearson correlations were as follows:

Speaking: 0.96 Listening: 0.97 Reading: 0.96 Writing: 0.93

5.2 Intra-rater reliability

Twelve months after the AMEP teachers had rated the 16 learners in Sydney, Melbourne and Adelaide, they were asked to re-rate them, using the same videotapes and scripts presented in a different random order in order to determine intra-rater reliability. The following is a summary of Spearman correlation coefficients:

Speaking

Correlations ranged from 0.90 to 0.99, with a mean of 0.97

Listening

Correlations ranged from 0.91 to 0.98, with a mean of 0.96

Reading

Correlations ranged from 0.87 to 0.99, with a mean of 0.97

Writing

Correlations ranged from 0.92 to 0.98, with a mean of 0.96.

In all cases, the level of significance was at or beyond 0.001. The data suggest a high level of intra-rater reliability (Ingram, 1984).

5.3 Test-retest reliability

The initial trialling addressed only the process of rating; that is, the teachers were making judgements about proficiency on the basis of language data elicited by other testers. When ISLPR tests are administered in the field – as they were in the studies by Phillips *et al* (1985), Lee (1992), Kellett and Cumming (1995), Gariano (1997), and Wylie (2001) referred to above – the tester is responsible for both eliciting and judging the data.

The practicum sessions that are an essential part of ISLPR training programs have provided test-retest data under rigorously controlled conditions. At the end of 20-hour ‘refresher’ training programs, the same volunteer learner is tested in speaking, listening and reading first by one trainee, using tasks that s/he has developed and selected for the particular learner, and then by another trainee, using different tasks. (For practical reasons, volunteers are not expected to do two writing tests.) There is no opportunity for collusion between trainees. Agreements on test-retest ratings on 84 learners tested by 84 trainees in 1994/5 were as follows:

Speaking: 0.90 Listening: 0.88 Reading: 0.89 (Wylie, 1996).

In all cases, the level of significance was at or beyond 0.001. These data suggest that candidates and other end-users of test results can be confident of the quality of ISLPR tests conducted by trained, experienced testers.¹¹

References

- American Council on the Teaching of Foreign Languages. 1986. *ACTFL proficiency guidelines*. Hastings-on-Hudson, NY: Author.
- Australian Foundation for the Peoples of Asia and the Pacific. 2000. Confidential report on a program to develop English language and computer skills of students in East Timor funded under AusAID’s Interim Capacity Building Program, East Timor (CAPET) and managed by The Illawarra Technology Corporation (ITC). Wollongong: ITC.
- Australian Language and Literacy Council. 1996. *Language teachers: the pivot of policy*. Canberra: National Board of Employment, Education and Training/AGPS.
- Beaumont, A. 1982/83. Materials lists for the Clearing House for the Adult Migrant Education Service. Adelaide: SA Migrant Education Service.
- Beaverson, A. and Carstensen, C. *Starters: Dialogues for beginning ESL students*. Brisbane: Division of Special Education, Department of Education, Queensland.
- Clark, J. and Swinton, S. 1979. An exploration of speaking proficiency measures in the TOEFL context. *TOEFL Research Report No.4*. Princeton: Educational Testing Service.
- Coppell, B., Baumgart, N. and Tenezakis, M. 1984. *English language learning needs of migrants in Parramatta*. Studies in adult migration (Smolicz, J., Series editor). Canberra: Department of Immigration and Ethnic Affairs/AGPS.
- Department of Immigration and Ethnic Affairs, Education Branch. 1979. *Methodology, proficiency ratings and language content*. A summary statement of developments relating in particular to the On-arrival Initial Settlement Education Program and resulting from the work of the Joint Commonwealth/States Committee on Professional Aspects of the Adult Migrant Education Program. Canberra: Author.
- Foreign Service Institute School of Language Studies. 1968. *Absolute Language Proficiency Ratings*. Washington, DC: Author.
- Gariano, A. 1997. *The role of English language competence on migrant settlement*. Unpublished doctoral thesis. Sydney: University of NSW.

¹¹ For high-stakes situations, however, systematic monitoring of testers’ methods and ratings and on-going feedback are essential. See Footnote 4.

- Gibbons, J. 1990. Applied linguistics in court. *Applied Linguistics*, 11(3): 229-238.
- Griffith University, 2000. Rationales for Graduate Diploma of Indonesian Language (Distance) and Graduate Certificate in Intermediate Indonesian (Distance). Brisbane: School of Languages and Linguistics, Griffith University.
- Ingram, D.E. 1979. Methodology. In *Adult Migrant Education Program Teachers Manual*. Canberra: Department of Immigration and Ethnic Affairs.
- Ingram, D.E. 1984. *Report on the formal trialling of the Australian Second Language Proficiency Ratings (ASLPR)*. Canberra: Department of Immigration and Ethnic Affairs.
- Ingram, D.E. 2010. *National Testing of English language skills for taxi drivers*. Invited paper to the Australian Taxi Industry Association (AITA) Conference, Darwin Convention Centre, Darwin, NT. 15 July.
- Ingram, D.E. and John, G. 1990. *The teaching of languages and cultures in Queensland: Towards a language education policy for Queensland schools*. Brisbane: Centre for Applied Linguistics and Languages, Griffith University.
- Ingram, D.E. and Lee, T. 1993. *Report on the profile of community proficiency in languages other than English conducted by AGB Australia for the Australian Language and Literacy Council*. Brisbane: Centre for Applied Linguistics and Languages, Griffith University.
- Ingram, D.E. and Wylie, E. 1979. Australian Second Language Proficiency Ratings (ASLPR). In *Adult Migrant Education Program Teachers Manual*. Canberra: Department of Immigration and Ethnic Affairs.
- Ingram, D.E. and Wylie, E. 1984. Australian Second Language Proficiency Ratings (ASLPR). Canberra: Department of Immigration and Ethnic Affairs.
- Ingram, D.E. and Wylie, E. 1991. Developing proficiency scales for communicative assessment. *Language and Language Education: Working papers of the National Languages Institute of Australia* 1(1):31-60.
- Ingram, D.E. and Wylie, E. 1996. Preliminary ASLPR – NRS (National Reporting System) Alignments. Brisbane: Centre for Applied Linguistics and Languages, Griffith University.
- Ingram, D.E., Lee, T. and Wylie, E. 1996. Proficiency: the ASLPR Chinese version. In Farquhar, M. and McKay, P. *China connections: Australian business needs and university language education*. Canberra: NLLIA.
- Interagency Language Roundtable. 1985. *Language Skill Level Descriptions*. Washington, DC: Author.
- Jensen, M. 1995. Linguistic evidence accepted in the case of a non-native speaker of English. In Eades, D. *Language in evidence: Issues confronting Aboriginal and multicultural Australia*. Sydney: UNSW Press.
- Joint Commonwealth-States Committee on Professional Aspects of the Adult Migrant Education Program. 1979. *Methodology, proficiency ratings and language: a summary statement of developments in the Adult Migrant Education Program*. Canberra: Education Branch, Department of Immigration and Ethnic Affairs
- Kellett, M. and Cumming, J. 1995. The influence of English language proficiency on the success of non-English speaking background students in a TAFE vocational course. *Australian and New Zealand Journal of Vocational Education Research* 3(1): 69-86.
- Lee, T. 1992. *A many-faceted Rasch analysis of ASLPR ratings*. Report to the Steering Committee for the Australian Assessment of Communicative English Skills Test. Brisbane: Centre for Applied Linguistics and Languages, Griffith University.
- Lee, T., Wylie, E. and Ingram, D.E. 1998. *Mapping rates of progress in proficiency*. Paper to the 20th International Language Testing Research Colloquium. Monterey, Ca., 9-12 March.
- Linacre, J. and Wright, B. 1990. *FACETS*. MESA Press.
- Manton, J., McKay, P. and Clyne, M. 1983. *English language learning needs of migrants in the suburbs of Melbourne*. Studies in Adult Migration (Smolicz, J., Series editor.). Canberra: Department of Immigration and Ethnic Affairs/AGPS.
- Matthews, E. 2004. *ICAO Language Proficiency Requirements*. Presentation at ICAO Regional Seminar on Language Proficiency. Tokyo, 8-10 Dec. Summary on ICAO website <http://www.icao.int/icao/en/anb/meetings/IALS/proceedings/PPTs/5-Matthews.pdf>, accessed August, 2008.

- Messick, S. 1994. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 23(2): 13-23.
- Mousavi, S. 2007. *Development and validation of a multi-media computer package for the assessment of oral proficiency of adult ESL learners: Implications for score comparability*. Unpublished doctoral thesis, Griffith University.
- National Languages and Literacy Institute of Australia (NLLIA). 1996. *Asian Languages Teacher Proficiency Project – Indonesian*. Final report to the Department of Employment, Education, Training and Youth Affairs. Canberra: Author.
- Oscarson, M. 1997. Self-assessment of foreign and second language proficiency. In Clapham, C. and Corson, D. (eds) *Encyclopedia of Language and Education, Vol 7, Language Testing and Assessment*.: 175-187.
- Phillips, D., Burke, E., Ingram, D.E. and Campbell, A. 1985. *The formative evaluation of preparatory English language training of sponsored Indonesian students*. Report to ADAB. Canberra: University of Canberra.
- Sefton, J. and Wylie, E. 2002. *Maximising the satisfaction of international students and minimising their attrition*. Report on a Quality Enhancement Grant. Griffith University (confidential).
- Sefton, R. and O'Hara, L. 1992. *Report of the work placed education project survey on behalf of the vehicle manufacturing industry*. Melbourne: Victorian Automotive Industry Training Board.
- Singh, P., Parker, K. and Dooley, K. 1999. *Technical workplace English for the Indonesian mining sector*. A report on a project funded by the Australian Research Council Strategic Partnership with Industry - Research and Training Scheme. Brisbane: School of Cognition, Language and Special Education, Griffith University.
- Smith, K. and Baldauf, R. 1982. The concurrent validity of self-rating with interviewer rating on the Australian Second Language Proficiency scale. *Educational and Psychological Measurement*, 42(4): 1117-1124.
- Tasmanian Educational Consortium. 1997. Education (LOTE Teaching). A Graduate Certificate for primary LOTE teachers. Hobart: Author.
- University of New South Wales Institute of Languages (UNSWIL). no date. *Professional English Assessment for Teachers (PEAT): Information for candidates and exemplar*. Sydney: UNSWIL and Department of Education and Training.
- Vygotsky, L. 1978. *Mind in society: the development of higher psychological processes*. Cambridge, Mass.: Harvard University Press.
- Willcock, H. 2001. Language majors – Japanese. School of Languages and Linguistics, Griffith University website <http://www.gu.edu.au/ua/aa/hbk/fsartsub5.html>, accessed September 2003.
- Wood, D., Bruner, J. and Ross, G. 1976. The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry* 17, 69-100.
- Wylie, E. 1996. In search of the perfect match: direct assessment of second language proficiency according to the ASLPR. *The Digest of Australian Language and Literacy Issues*, 14. Canberra: Language Australia.
- Wylie, E. 1997. Test and non-test assessment for the professional second language user. In Huhta, A., Kohonen, V., Kurki-Suonio, L. and Luoma, S. (eds) *Current developments and alternatives in language assessment*. Jyväskylä, Finland: University of Jyväskylä.
- Wylie, E. 2000. Proficiency scales as a means of integrating the teaching/learning curriculum and assessment. Paper presented at the English Australia conference held at the Esplanade Hotel, Fremantle 12 to 14 October. In *Proceedings of the 13th Annual EA Education Conference*. Sydney: English Australia, 245-252.
- Wylie, E. 2001. *Issues in self-assessment of proficiency in a foreign language*. Unpublished paper submitted as part of requirements for a doctorate in education, Griffith University.
- Wylie, E. and Ingram, D.E. 1995/99. *International Second Language Proficiency Ratings (ISLPR): General proficiency version for English*. Brisbane: Centre for Applied Linguistics and Languages, Griffith University. Revised 2010, ISLPR Language Services, Brisbane.
- Wylie, E. and Ingram, D.E. 2010. *How good should a taxi driver's English be?* Paper to the Language Testing Centre 20th Anniversary Event, University of Melbourne, 15 July.