Scales and tests—competition or cooperation?

Kathryn Hill

Alderson (1991) proposes a number of ways in which language proficiency scales may relate to language tests. However, in the current Australian context, the two are widely seen to represent distinct, if not competing, approaches to language assessment. This paper considers the relationship between scales and tests in the context of a language proficiency assessment project which aims to incorporate both elements. The paper begins with an attempt to define the essential characteristics of standardised language tests before going on to describe the Indonesian teacher proficiency project and the method by which it is intended scales and tests might be used cooperatively.

1. Language tests

A language test may be defined as a language task with known measurement properties. The task may vary in type (multiple choice, note taking, cloze etc.) and level of difficulty. It may aim to measure a single skill (e.g. distinguishing between minimal pairs), integrated skills (e.g. writing an essay) or interactive skills (e.g. participating in a discussion). The test may be used to assess achievement on a course of study, for diagnosis, selection or certification, or for a combination of purposes.

However, whilst they may take a number of forms and fulfil a multiplicity of purposes, the essential qualities demanded of language tests remain constant. The following represents an attempt to briefly specify these qualities.

One claim that is made for language tests is that they provide an operational definition of proficiency. That is, a test aims to tap a specific body of language knowledge and abilities and, to the extent that it succeeds, candidates' proficiency is defined by performance on the test. It is therefore important that test designers , firstly, be explicit about what is to be measured and, secondly, ensure that the test is a valid measure of that ability. Types of evidence for test validity include content validity, construct validity and criterion-related validity, each of which is briefly described below.

Melbourne Papers in Language Testing 1995 Volume 4.2 pp. 43-59.
The Language Testing Research Centre, The University of Melbourne.

Content validity depends, initially, upon a thorough linguistic analysis of the target domain (e.g. English for health professionals) and, subsequently, upon principled sampling from the language domain so defined. To ensure adequate content coverage, the advice of a range of 'experts' (e.g. subject specialists, industry representatives, etc.) is commonly sought in the early stages of test development . However, once the test has been developed and trialed, some sort of attempt needs to be made to determine whether test items actually measure the skills and abilities they were originally designed to measure, ora posteriori content validation (Alderson 1988).

Construct validity, by one definition, involves examination of the integrity of the model of language, or construct, which has provided the rationale for the test. More commonly, however, construct validation entails an attempt to identify those skills or abilities the test is actually measuring (which may or may not be the same as what it set out to measure). Methods employed here include correlational studies, which use factor analysis to identify the number of factors in the test data (e.g. Multi-Trait Multi-Method, Campbell & Fiske 1959), investigation of test-taking strategies (e.g. Cohen 1988, Cohen & Hosenfeld 1981) and 'cross-sectional' studies comparing the performance of groups which might be expected to differ on the construct being measured, for example, comparisons between experts and novices (Messick 1988: 55). Finally, feedback from test participants (i.e. test takers, interviewers, raters, etc.) may also be used to provide additional evidence for construct validity (e.g. Brown 1991, Hill 1994).

Predictions are often made on the basis of test performance as to how the candidate would perform in some future context. Validity, therefore, does not reside exclusively in the specifications and the resulting test items but 'in the relation between the test and the domain of application', or 'criterion' (Messick 1988: 41). This criterion may be performance on another, comparable, test (concurrent validity) or it may be future, real-world performance (predictive validity). In practice, predictive validation studies, for example, those comparing performance on English for Academic Purposes tests and academic performance, tend to find relatively weak relationships between test and real-world performance (e.g.

Criper & Davies 1988, Elder 1993)[1]. Nonetheless, prediction of criterion performance remains fundamental to test validation.

Finally, validity demands that a test be reliable as well as linguistically meaningful. The issue of reliability is particularly important for 'high stakes' assessment (e.g. where a test is used for selection or certification). Test reliability is affected by a number of factors which need to be understood and controlled in order to reduce measurement error. Potential sources of error include the test items, test method, test length, the people making the assessment and the conditions of test taking, each of which will be discussed in turn.

(i) Test items

Information about candidates' ability is gained by means of their performance on a specified task. As it is not always possible to predict how candidates will respond to a given task, tasks first need to be trialed on a proper sample from the relevant population. Trialing provides important information about the properties of the test including whether it is at an appropriate level of difficulty and whether it discriminates reliably between candidates. If alternative tasks are used, they need to be formally equated (both in terms of the abilities measured and in terms of difficulty) so that candidates are not disadvantaged by attempting one task rather than another.

(ii) Test method

Test designers also need to ensure that the method used to test candidates does not interfere with demonstration of the language ability in question. Research suggests that lower proficiency candidates are more sensitive to test method than more proficient candidates (e.g. Shohamy & Inbar 1991). To help neutralise any possible 'method effect', it is often recommended that tests comprise a range of different item types.

[1] This result should not be surprising considering that academic success is attributable to many factors in addition to language ability.

(iii) Test length

A test needs to elicit an adequate sample upon which to base a judgement about a candidate's ability. Generally speaking, the longer the test, the more information provided about the candidate's proficiency and the higher the reliability. This is why well-established international tests typically take over two hours to administer.

(iv) Judgement

Any test item elicits a performance from the candidate and the score awarded

is derived from a judgement of that performance. For objectively scored tests, where there is only one possible answer (e.g. multiple choice, true/false etc.), judgement is not an issue. However, for subjectively scored tests (e.g. tests of speaking and writing), consistency of judgement within and between raters becomes more of a concern. For this reason, the training of raters and moderation of assessments is essential. As multiple judgements improve the reliability of assessment, subjectively rated tests should be routinely double marked, with discrepancy marking in the case of significant disagreement. A computer program is available which is able to compensate for variability in rater severity when estimating candidate ability (FACETS, Linacre 1989-92) but, once again, this program depends upon multiple ratings for each candidate.

(v) Test taking conditions

The are a number of additional factors which have been shown to influence test performance and, hence, the accuracy of the measurement. Such conditions include the test setting, (i.e. location and time of testing), test rubric (i.e. instructions, layout, time allowances) and mode of assessment (e.g. oral interview, audio-tape, video-tape).

All of these potential sources of measurement error need to be understood and their effects controlled or minimized in order to ensure candidates are treated fairly and measured as accurately as possible.

2. Scales and tests

Scales[2] and tests are widely seen to represent two distinct, if not competing, approaches to language assessment in Australia. The question to be considered here, therefore, is how the two approaches might be used cooperatively.

The Asian Languages Teacher Proficiency Project - Indonesian (ALTPP), a project to develop assessment procedures for teachers of Indonesian in Australia, represents a deliberate attempt to reconcile the two approaches. The project is being carried out jointly by the NLLIA-Language Testing Research Centre (LTRC) at the University of Melbourne and the NLLIA-Language Testing and Curriculum Centre (LTACC), at Griffith University. LTACC is responsible for the development of a set of proficiency scales (herein referred to as 'the scales') and the LTRC is responsible for test development.

The scales are a specific purpose version of the Australian Second Language Proficiency Ratings (ASLPR) for teachers of Indonesian (Wylie, Ingram & Woollams 1995). They comprise nine levels of proficiency (from "zero" to "native-like" competence) and provide a one page description for each level of each macro skill[3]. These scales are to be used to describe Indonesian language proficiency as measured by standardised tests developed for each of the four macro skills (described in greater detail below).

To understand the reason for this approach, it is necessary to provide some idea of the background to the project. The Indonesian teacher proficiency scales are an adaptation of a specific purpose version of the ASLPR for second language teachers (i.e. not specific to any language). These scales were produced as part of an earlier project set up to develop descriptions (or prescriptions)[4] of minimum

[2] The characteristics of scales are discussed in detail in Ingram's article (same volume) and, as such, will not be elaborated here.

[3] These descriptions exist for only six of the nine levels.

[4] As they were developed in consultation with an Advisory Committee, rather than on the basis of empirical data on language teacher proficiency, I would

competencies for language teachers as part of a national strategy to upgrade the status of language teaching in Australia. The current project forms the next step in the process; devising a means of assessing language teacher proficiency against these descriptions.

A point that is often not properly understood (in Australia, at least) is that scales are not tests: they may "provide a means of describing levels of proficiency, [but] do not in themselves measure that proficiency."(Rudd 140). What this means is that, for assessment purposes, the scales need to be used in conjunction with a language test or task; the test measures language ability and the scales describe language ability as measured by the test.

The ASLPR is usually used in conjunction with a bank of assessment tasks (typically assessor-generated), rather than standardised tests. Testing takes place in an oral interview (except writing) and is 'adaptive', i.e. each candidate may be presented with a different set of tasks, depending on his perceived level of proficiency. However, as pointed out earlier, where results may be used for selection or certification, it is important that assessment to be consistent and reliable. Such consistency is difficult to achieve when the content and duration of assessment differs for each candidate. The rationale for using standardised tests rather than 'orthodox' ASLPR assessment methodology, therefore, is that every candidate, weak or strong, is offered the same testing experience.

Whilst it can be argued that both components are necessary, the issue of how to ensure their compatibility remains. Alderson (1991) proposes three types of relationship which may exist between scales and tests. According to Alderson, scales may provide:

1. a guide to test design (constructor-oriented);

2. a guide for assessing performance on a test (assessor-oriented); and

3. a format for reporting performance on a test (user-oriented).

All three types of relationship could be said to operate in the Indonesian proficiency project and, for this reason, Alderson's

argue that the scales prescribe ideal levels of performance rather than describe actual levels of proficiency.

taxonomy has been used as a framework for describing the test development process. However, whereas Alderson's article seems to suggests a one-way influence, i.e. of scales on tests, the experience of the current project is that the relationship is, of necessity, interactive.

2.1. Test design

Because test performance will ultimately be reported as a level on the scales, it was considered necessary to ensure that test content was consistent with scale content. However, as Brindley (1995) suggests, such scales are too vaguely defined to translate directly into specifications. Rather, the main source of input for the specifications came from previous research at LTRC, specifically, the development of teacher proficiency tests for Italian (Elder 1993) and Japanese (Elder et al 1994).

What the Italian and Japanese teacher proficiency tests and the scales have in common is that each attempts to define the same language domain, i.e. the type of language proficiency considered necessary for language teaching. However, the realization of this domain, in the scales and tests respectively, results in less than a perfect fit either in terms of format or content. As a result, a number of compromises have proved necessary for the current project.

Listening/Reading

A range of texts was identified by NLLIA-LTRC and, following 'orthodox' ASLPR assessment practice, an attempt was made to 'match' them to a level on the scales (i.e. identifying each as a Level 1 text, a Level 2 text, and so on). This task was accomplished in conjunction with the LTACC language consultant (also an ASLPR assessor) and an Indonesian specialist from the Advisory Committee. For ASLPR-type assessment, the usual procedure is to place candidates on a scale level according to whether or not they can 'do' the text assigned to that level. For the Indonesian project, however, a range of items was devised for each text, with the result that a 'level 1' (i.e. 'easy') text may have some relatively difficult items attached to it and a 'level 5' (i.e. 'difficult') text some easy items. As it was generally agreed to be an essential skill for teachers, a question-writing task was also included on the trial

Reading test (despite LTACC's concerns about how and where to report this skill on the scales).

Writing

For the writing sub-test, two 'authentic' tasks were devised, a formal and an informal letter. Again, 'orthodox' ASLPR assessment practice would require a separate writing task to test each level on the scale. However, it was eventually conceded that this approach would make the Writing sub-test unreasonably lengthy.

Speaking

The Advisory Committee did not support the idea of using a paired interview format (used in the Japanese test) for the Speaking sub-test, and initially favoured the idea of an Oral Interaction test rather than separate sub-tests for Listening and Speaking. The latter proposal, however, proved unacceptable to LTACC as the scales require "unambiguous data" for each macro-skill. The trial version of the Speaking sub-test includes a range of task types, all of which are contextualised in the classroom (Appendix 1). In response to LTACC's concern that the stronger candidates may not be sufficiently challenged, a task involving an open-ended discussion was added.

A test of formal language knowledge, in the form of a written editing task, used in both the Italian and Japanese proficiency tests, was not included. The original objection was that there was no place in the scales for reporting on this ability. However, in the end an editing task was included in the Speaking sub-test and the scales amended to report on the ability to use Indonesian to talk about the language.

In summary, the development of the trial version of the Indonesian test was influenced by a number of factors. These included previous test development, the format and content of the scales, and, not least of all, the views of the project Advisory Committee.

2.2. Assessment

Writing and Speaking

Rather than use a separate task to assess each level of the scale (i.e. using 'orthodox' ASLPR assessment methodology) the approach taken for scoring the writing and speaking sub-tests was to define different levels of performance within each of the tasks. Initially, it was intended that the number of levels for each assessment category on the rating scheme would match the six described levels of the ASLPR scale. However, in practice, some of the levels were found to be too broad to distinguish properly between candidates who clearly differed in ability. It was then decided that, as ASLPR 0 (zero proficiency) and ASLPR 4 & 5 (native and virtual native-like proficiency) were unlikely to be awarded, these levels should be used to provide nominal top and bottom extremes of the scale. The result was a seven point rating scale (i.e. 0,1,2,3,4,5,6) with five 'usable' levels (i.e. 1,2,3,4,5).

Whilst the assessment categories (and accompanying descriptors) used to assess the writing and speaking tasks should be consistent with what is reported in the scales, ultimately they need to be appropriate to the task in question. That is, the assessment categories chosen need to come out of actual test performance and cannot necessarily be prescribed by the scales. In fact, the scale descriptors are very detailed and large sections of their content irrelevant to the writing tasks in question (i.e. a formal and an informal letter). It was, therefore, considered necessary to develop more 'user-friendly' rating schemes.

Draft descriptors were developed for the Writing sub-test using the same assessment categories as for the writing component (also a letter) of the Italian teacher proficiency test (i.e. Fluency, Content, Form & Overall). As appropriate, these descriptors also incorporated language features included under 'Descriptions of Language Behaviour' for each level of the scales. A language consultant was then asked to examine a range of scripts and make notes about the salient features of each performance. Each script was then discussed by the author and the consultant in relation to the rating scheme and the descriptors refined as necessary[5]. During trialing each writing task was double marked on the four assessment categories, resulting in a total of 16 scores per candidate. At the same time the LTACC language consultant was asked to independently assign each of these scripts to a level on the ASLPR scale[6].

A similar process was followed for marking the Speaking section.

2.3. Reporting

Reading and Listening

Test data were analysed using Quest (Adams & Khoo 1992), a program which maps candidate ability and item difficulty onto the same (logit) scale. When item difficulty was examined to see where the different texts and items fell along the scale, no hierarchy was evident for items associated with individual texts (Appendix 2). However, after an attempt was made to define what each item might be measuring, some clustering seemed to be evident for items measuring particular types of skills. On this basis it was possible to distinguish five different ability levels and to write level descriptors characterising the types of skills candidates demonstrated within each level. Again, care was taken to remain consistent with the content of the ASLPR scales during this process.

The level descriptors described above were developed solely for the purpose of reporting to trial candidates. However, in practice, test performance is to be reported as a level on the scales. To enable this to happen, a group of approximately 50 of the trial candidates will be assessed using both the test and orthodox ASLPR assessment methodology. A simple regression will then be performed to determine how a level on the scale can be calculated from a score on the test.

[5] These descriptors were further refined during rater training.

[6] Unfortunately, the consultant was not prepared to articulate which features of performance led him to assign a particular level to a particular candidate.

Regardless of the success of this 'benchmarking' exercise, it is essential that the scale description resulting from a test score is meaningful in terms of the original performance, i.e. that the candidate actually demonstrate the abilities attributed to her by the scales. As Brindley (forthcoming) suggests, this will be easier to establish for speaking and writing than for reading and listening, which are not observable[7].

Without empirical data on teacher proficiency (such as that provided by testing), the scales merely represent a prescription (cf description) of ideal levels of performance. If a close correspondence is found between the scale descriptions and test performance, this may be seen to provide evidence for the validity of both the scale (as a description of levels of language teacher proficiency) and the test (as a measure of that proficiency). If, on the other hand, it does not prove possible to relate test performance to what is reported in the scale, this may be attributed to a lack of comparability between two different approaches to assessment, each making different contributions to our understanding of language proficiency.

[7] A study has been initiated to determine whether what is measured by the Reading sub-test is consistent with what is reported by the scales. A group of Indonesian language specialists were asked to: classify each reading text according to type and linguistic features; rank the five texts in terms of difficulty, and identify the most critical skill required to answer each item.

Bibliography

Adams, R.J. & S.T. Khoo (1992) *QUEST*. Hawthorn, Victoria: Australian Council for Educational Research.

Alderson, J.C. (1988) 'New procedures for validating proficiency tests of ESP? Theory and practice.' *Language Testing* 5,2, 220–232.

Alderson, J.C. (1991) 'Bands and scores'. In J.C. Alderson and B. North (eds.), *Language Testing in the 1990s*. London: Macmillan, 71–86.

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

Brindley, G. (forthcoming) to appear in L.F. Bachman and A.D. Cohen (eds) *Interfaces between SLA and Language Testing Research*. Cambridge: Cambridge University Press.

Brown (1991) *The role of test-taker feedback in the validation of a language proficiency test*. Unpublished MA thesis, University of Melbourne.

Cohen, A.D. (1988) 'The use of verbal report data for a better understanding of test taking processes'. *Australian Review of Applied Linguistics*, 11: 30–42.

Cohen, A.D. & Hosenfeld (1981) 'Some uses of mentalistic data in second language research'. *Language Learning*. 31,2: 285–313.

Criper, C. & A. Davies (1988) *ELTS Validation Report London*: The British Council/Cambridge: University of Cambridge Local Examinations Syndicate.

Campbell, J.T. & D.W. Fiske (1959) 'Convergent and discriminant validation by the multitrait-multimethod matrix'. *Psychological Bulletin* 56: 81–105.

Elder, C. (1993) *The Proficiency Test for Language Teachers: Italian, Volume 1: Final Report on the Test Development Process*. National Project Report for DEET, Canberra. 34pp.

Elder, C. (1993) 'Language proficiency as a predictor of performance in teacher education'. *Melbourne Papers in Language Testing*.2,1, June 1993, 68–85.

Elder, C. (1994) 'Performance testing as a benchmark for LOTE teacher education' *Melbourne Papers in Language Testing*.3,1, May 1994, 1–25.

Elder, C., N. Iwashita & A. Brown (1994) *The proficiency test for language teachers: Japanese.* Final report submitted to DEET, Canberra.

Hill, K. (1994) *The contribution of feedback to the development and validation of an oral proficiency test.* Unpublished MA thesis, University of Melbourne.

Linacre, J.M. (1989–92) *FACETS, a computer program for the analysis of multi-faceted data.* Chicago: MESA Press.

Lumley, T. (1993) 'Reading comprehension sub-skills: teachers' perceptions of content in an EAP test'. *Melbourne Papers in Language Testing.* 2,1, June 1993.

Messick, S. (1988) 'Validity'. In Linn, R.L. (ed.) *Educational Measurement.* Third Edition, N.Y.: Americal Council on Education/Macmillan.

Shohamy E. & O. Inbar (1991). 'Valididation of listening comprehension tests: the effect of text and question type'. *Language Testing*, 8, 1: 23–40.

Wigglesworth, G. & K. O'Loughlin (1993) 'An investigation into the comparability of direct and semi-direct versions of an oral interaction test in English'. *Melbourne Papers in Language Testing.* 2,1.

Wylie, E., D. Ingram & G. Woollams (1996) *ASLPR Version for Indonesian teachers.*

Appendix 1. Indonesian speaking test (trial version)

Phase 1: Warm-up

Candidates engage in a brief "getting to know you" conversation with the interviewer. The main purpose is to put the candidate at ease.

Target audience: interviewer

Type of communication: "extra-classroom use" Mode: Dialogue

Functions: requesting/giving information/expressing opinions etc.

Phase 2: Reading aloud

The candidate reads aloud a short text as if to a group of students and then, at the request of the interviewer, explains *in Indonesian,* the meaning of selected words from the passage

Target audience: whole class

Type of communication: "message-oriented" Mode: Monologue

Functions: narrating, explaining, exemplifying

Phase 3: Giving Instructions

Using a set of picture prompts the candidate explains, as if to a group of L2 learners, how to participate in a language game.

Target audience: whole class

Type of communication: "activity-oriented" Mode: monologue

Functions: directing, explaining

Phase 4: Modelling a role play

Candidate and interviewer then act out an authentic situation together.

Target audience: individual (native speaking) student/parent etc.

Type of communication: "medium-oriented" Mode: dialogue

Functions: explaining, persuading, requesting information, complaining etc.

Phase 5: Explaining learner error

Using an authentic piece of student writing in Indonesian, and in response to prompting from the interviewer, the candidate explains, as if to a second language learner, the nature of his/her mistakes.

Target audience: individual student

Type of communication: "medium-oriented" Mode: dialogue

Functions: explaining/ eliciting information/use of Indonesian metalanguage

Phase 6: Discussion

Interviewer engages candidate in an extended discussion using the topic of reading passage (Phase 2) as the starting point.

Target audience: interviewer

Type of communication: "extra-classroom use" Mode: dialogue

Functions: requesting/giving information, expressing opinion, speculating

Appendix 2. Item (text+task) difficulty compared with candidate ability

```
--------------------------------------------------------------------------------
Item Estimates (Thresholds)                                 10-Apr-96 17:19:01
all on all (N = 143 L = 53 Probability Level= .50)
--------------------------------------------------------------------------------
                                   |    Txt4
                                 X |
                                   |
 4.0                          XXXX |
                                   |
                                 X |
                                   |
                               XXX |
                                   |
 3.0                      XXXXXXXX |    Txt5
                                   |    Txt5
                          XXXXXXXX |
                            XXXXXX |    Txt4
                               XXX |
                                   |
                             XXXXX |    Txt1
 2.0                        XXXXXX |    Txt4
                               XXX |    Txt2   Txt2   Txt4 Txt5
                              XXXX |    Txt5   Txt5
                              XXXX |
                    XXXXXXXXXXXXXX |
                                 X |    Txt5
 1.0                       XXXXXXX |    Txt1
                              XXXX |    Txt5   Txt5
                          XXXXXXXX |    Txt5   Txt5
                                XX |    Txt1   Txt3
                                XX |    Txt3   Txt5   Txt5
                        XXXXXXXXXX |
                              XXXX |    Txt4   Txt4
  .0                          XXXX |    Txt4
                             XXXXX |    Txt2   Txt3
                              XXXX |    Txt5
                             XXXXX |    Txt2   Txt4   Txt5
                              XXXX |    Txt2
                             XXXXX |    Txt4
                                   |    Txt1   Txt3   Txt4   Txt4
-1.0                             X |    Txt2   Txt3   Txt3   Txt5
                                XX |    Txt3   Txt5
                                XX |    Txt2   Txt3   Txt3
                                 X |
                                   |    Txt3   Txt3
                                 X |    Txt3
                                   |    Txt4
-2.0                               |    Txt2
                                   |    Txt4
                                 X |    Txt1
                                   |    Txt2
                                   |    Txt3
-3.0                               |    Txt1
--------------------------------------------------------------------------------
 Each X represents    1 student
================================================================================
```