Scales Of Language Proficiency[1]

Brian North

1.0 Introduction

Scales of language proficiency have become relatively widespread over the past ten years or so as part of a general movement towards more transparency in educational systems and as greater international integration—particularly in Europe—places a higher value on being able to state what the attainment of a given level of language proficiency means in practice. Whereas 10 or 15 years ago, scales which were not directly or indirectly related back to the 1950s US Foreign Service Institute (FSI) scale (Wild 1965) were quite rare, the last few years has seen quite a proliferation with, for example, the British National Language Standards (Languages Lead Body 1992); the Finnish Scale of Language Proficiency (Luoma 1993) and the ALTE Framework (Association of Language Testers in Europe 1994). Many of these scales (see pages 143–44 for a list of those used as sources in this project) represent what Bachman (1990: 325–330) has described as the "real-life" or behavioural approach to assessment in that they try to give a picture of what a learner at a particular level of attainment can do in the real world. Other scales take what Bachman describes as the "interactive-ability" approach focusing upon aspects of a performance in a particular test (e.g. Milanovic et al 1992; Fulcher 1993; Upshur and Turner 1995; Brindley forthcoming). The following extract from the mid range of the 10 band Eurocentres global scale is a fairly typical yet simple example of the "real life" approach. This scale, it should be stressed, is the pinnacle of an information pyramid with more detailed scales used for the different purposes.

[1]This article is contains Chapters 2 & 3 of the author's PhD thesis "The Development of a Common Framework Scale of Language Proficiency Based on a Theory of Measurement." Thames Valley University, 1996.

Melbourne Papers in Language Testing 1995 Volume 4.2 pp. 60-111. The Language Testing Research Centre, The University of Melbourne.

7	Can express ideas and opinions clearly on a wide range of topics, and understand and exchange information reliably. Has an active command of the essentials of the language. Can communicate competently and independently in many professional as well as personal contexts.
6	Can understand information on topics of interest in unsimplified but straightforward language and can find different ways of formulating what he or she wants to express. Has assimilated the essentials of the language. Can communicate competently in many professional as well as personal contexts.
5	Can understand extensive simple information encountered in everyday situations and maintain conversation and discussion on topics of interest. Can exploit a wide range of simple language flexibly to express much of what he or she wants to. Can communicate adequately in routine professional contexts.

The style of this particular scale is deliberately simple. It is intended to give meaning to the numbers at a very general level, primarily to help students orient their learning. A purely numerical scale like the TOEFL scale can mean quite a lot to insiders, but does not say much to someone unfamiliar with TOEFL.

That this particular scale does fulfil its purpose is suggested by a recent study in connection with a joint project between the Swiss Bureau for Trade and Industry and Eurocentres in which approximately 120 young long term unemployed were sent on a 3 month stay abroad on Eurocentres courses in autumn 1994 to see if this would improve their employability through the acquisition of increased self esteem and better language skills. Before the stay, each learner's starting position on the scale was determined in order to agree a learning contract. Each learner took a short written test of knowledge of the language system—drawn from a validated item bank for English (Jones 1993)—and a formal interview with rating onto the Eurocentres assessment grid (usually used for rating classroom group interaction: North 1991; 1993b). The written tests were provided with a transformation table converting results onto the Eurocentres scale (the one for English having been derived empirically) and the average of these two tests was used to place the learners on the scale. On the same occasion, before taking the tests, the learners were also asked to read the scale in their mother tongue and assess their position on it. The correlation between the self assessments and the placement on the scale deduced from the combined test scores was 0.74 (n=104; p = .001) for English, French

and German taken together, and 0.78 for English taken alone (n = 58; p = .001). Of the learners of English, 43% rated themselves onto exactly the same scale band as the combined tests. For French and German, taken together, this proportion was only 20%, probably due to the fact that the assessment instruments themselves were still in the process of being validated. The magnitude of the correlations reported above are almost exactly the same as the correlation achieved on repeated occasions between such global test placement and Cambridge examinations (North 1991; 1994) and compare favourably with the sorts of correlations between self assessment and tests commonly reported in the literature (See Oscarson 1984 & Blanche 1986 for reviews and e.g. Blanche 1990; Swain 1992; Smith 1992; Wesche et al 1993).

The kind of transparency that this scale apparently had for those learners is the advantage that scales defining bands of language proficiency have over test scores or numerical scales (e.g. 1–1,000) and is one reason why they are becoming more and more popular.

1.1 Definitions

Scales of Language Proficiency go by many different names, for example "band scores, band scales, profile bands, proficiency levels, proficiency scales, proficiency ratings" (Alderson 1991a:71) or "guidelines, standards, levels, yardsticks, stages, scales, or grades" (De Jong 1992:43). What they all have in common is that they attempt to provide "an ascending series of levels of language competence" (Page in North et al 1992:7) or "a hierarchy of global characterisations of integrated performance" (ACTFL 1986), "a hierarchical sequence of performance ranges" (Galloway 1987:27) or "characteristic profiles of the kinds and levels of performance which can be expected of representative learners at different stages" (Trim 1978:6).

Definitions of scales of language proficiency in the literature depend somewhat on the perspective of the writer and the argument they are in the process of putting forward. John Clark's definition catches their main weakness: "descriptions of expected outcomes, or impressionistic etchings of what proficiency might look like as one moves through hypothetical points or levels on a developmental continuum" (Clark 1985:348). In other words, scales of language proficiency give pictures of successive levels of language learning

attainment, and although users may well be able to interpret them with some success (as in the case given above), this doesn't necessarily mean that what the scales say is actually a valid description of stages of the second language acquisition process since "the generalised descriptions of levels which figure in rating scales represent an inevitable and possibly misleading oversimplification of the language learning process" (Brindley forthcoming: 22).

1.2 Attractions

Nevertheless, despite Clark's and Brindley's reservations, scales of proficiency have been noted to offer a number of attractions. They can be used to:

- provide a "stereotype" with which the learner can compare his self image and roughly evaluate his position—as in the case cited at the beginning of this chapter (Trim 1978; Oscarson 1978, 1984);

- establish a framework of reference which can describe achievement in a complex system in terms meaningful to all the different partners in or users of that system in a way that scores from test items cannot (Trim 1978; Brindley 1986 1991; Schneider and Richterich 1992);

- provide learner goals and descriptions of proficiency at notional levels in order to provide targets for learners, to allow the results achieved to be measured against expected outcomes, and to provide society with a pragmatic means of placing students in appropriate future learning or work environments by referring to an individual's profile across the sub-scales of the system (Clark 1985);

- provide coherent internal links within one system between pre-course or entry testing, syllabus planning, materials organisation, progress and exit assessment and certification (North 1991);

- provide evidence of progress (provided the steps are small enough) and so help increase motivation (Liskin-Gasparro 1984a, Page 1992; North 1992a);

- increase the reliability of subjectively judged ratings, especially of the productive language skills, and provide a common standard and meaning for such judgements (Alderson 1991a);

- report results from teacher assessments, scored tests, rated tests and self assessment all in terms of the same instrument and avoid the spurious suggestion of precision given by a scores scale (e.g. 1–1,000) (Alderson 1991a; Griffin 1989);

- provide achievement stages and grades which reflect the curriculum of the classroom, but which can be translated into a proficiency statement and grade on a common framework (Trim 1978; Ingram & Wylie 1989; Hargreaves 1992);

- enable comparison between systems or populations using a common metric or yardstick (Trim 1978, Lowe 1983, Liskin-Gasparro 1984b; Bachman and Savignon 1986; Carroll B.J. and West 1989);

Thus, despite the problems attached to trying to describe complex phenomena in a small number of words on the basis of incomplete theory, which was referred to at the end of the previous section, scales of language proficiency have the potential to exert a positive influence on the orientation, organisation and reporting of language learning.

1.3. Origins

The many names given to scales of language proficiency mentioned in Section 1.1. reflect the fact that such scales can have quite different backgrounds. Scales of language proficiency seem to have one of three types of origin:

- as rating scales;

- as examination levels;

- as stages of attainment;

1.3.1. Rating Scales:

The majority of existing scales of language proficiency are in effect rating scales which have holistic definitions attached to the steps on the scale. They are scales for assigning a grade in a test to which descriptions have been added for each level, and which have gone on to acquire a "framework" role as people have started using them as a point of reference. Ultimately, scales of proficiency are derived from the items which are found on opinion polls or questionnaires, which are also referred to as scales (Skehan 1989a:10–12). What has happened is that the scales have been turned vertically, a behavioural definition has been given to each category instead of or in addition to the original value label like "Good" (Champney 1941), and different dimensions have been separated and presented on different pages (Smith and Kendall 1963) or in the columns of a grid to produce an analytic as opposed to holistic scale (Shohamy 1981).

The first significant scale of language proficiency was the rating scale of the US Foreign Service Institute, (the FSI scale) developed in the early 1950s. The FSI is the direct forerunner of the ASLPR (Australian Second Language Proficiency Ratings), the ILR (Interagency Language Roundtable) scale for US government employees and the ACTFL (American Council of the Teaching of Foreign Languages) Proficiency Guidelines. The first three share the same scale bands; ACTFL have developed narrower bands in the lower part of the scale, but claim equivalence with ILR through their common origin.

The FSI scale had six steps from zero (Foreign) to perfection (Native: the now notorious "educated native speaker" or ENS) and raters judged relative amounts of foreignness or nativeness of each so-called "factor": accent, fluency, comprehension, vocabulary and grammar" (Lowe 1985:19). The criterion for this judgement was the set of holistic descriptions of performance for Speaking and for Reading for each of the six levels—which we know as the FSI scale—which had been elaborated from a set of descriptors prepared for a 1955 survey of foreign language skills in the Foreign Service department (Liskin-Gasparro 1984b:18–19).

In the case of the FSI scale, then, a reporting framework or set of behavioural descriptors developed in tandem with the test used to situate people on it—the FSI oral interview. This test had no "pass", no criterion-score. As with the ILR, ACTFL and ASLPR

which are all developed from it, candidates were situated in an ascending series of levels covering the continuum of language proficiency. Many other tests, especially oral tests, have also developed such scales of descriptors (See e.g. Carroll 1980; Shohamy 1981; Morrow 1977; 1986; Alderson 1991; Milanovic et al 1992).

It is, however, unlikely that any scale of language proficiency has been developed without being directly or indirectly influenced by the FSI approach. "Like tests, some proficiency scales seem to have acquired popular validation by virtue of their longevity and extracts from them appear regularly in other scales" and it is very difficult to find out how particular descriptors were arrived at (Brindley 1991:6–8). This leads to two main potential problems:

- Descriptors which may have been appropriate for use by a particular group of raters for a particular purpose in one particular context may be picked up and used for a range of different purposes with different populations (c.f. Spolsky 1986:148 discussing the development of the ACTFL guidelines from the FSI scale). Validity can only be seen as relative to function and context: "What is this test/scale valid for?" rather than "Is this test/scale valid?" (Henning 1990:379).

- Decisions about what level to put particular tasks may be purely the result of convention and the task descriptors may be just clichés which get copied from scale to scale (North 1992a:168). Since a main purpose of descriptors is to "anchor" judgements—as in "Behaviourally Anchored Rating Scales—BARS—the effect of conventions and clichés not based on any empirical evidence may be to systematise the very judgement error the definitions are intended to help avoid (Landy and Farr 1983).

One should note that not all rating scales have developed into scales of language proficiency in the sense in which it is being used in this thesis. In many examinations, a rating scale is normed around the pass level with minimal and heavily relative wording. For example in the Cambridge First Certificate Oral Interview (Paper 5) a rating scale is used to grade the candidate in relation to the pass norm on a set of 6 factors or criteria defined with a sentence for each level. It seems to be a feature of such rating scales that you often need to be virtually a native speaker to get the top grade (a 5

in this case) despite the fact that the rating scale is only relevant to one examination in a series (in this case the second exam out of four). Since the scale is interpreted only in relation to the examination norm, the wording of the rating scales for different examinations in the suite may even be extremely similar even though the performance expected may be very different. For a very long time the Cambridge First Certificate and Certificate of Proficiency examinations used norm-referenced rating scales for their oral papers which were virtually identical, yet appear to have meant different things in the context of the two examinations.

1.3.2. Examination Levels:

Suites of examinations can, however, contribute to the development of scales of language proficiency in a different way. The suite of communicative examinations developed by the Royal Society of Arts (RSA) for English as a Foreign Language offers a classic example of how this can happen. The suite of exams (now administered by Cambridge: (University of Cambridge/Royal Society of Arts 1990) were originally developed following the recommendations of Morrow (1977). The exams have defined content and performance specifications for each level, the categories for the performance criteria—called "degrees of skill"—remaining the same for each of the 4 levels. The criteria for Oral Interaction, for example, are: Accuracy, Appropriacy, Range, Flexibility and Size. Although rating is on a simple pass/fail mastery rather than scalar basis, with assessors matching the candidates performance to the definitions of the 5 criteria for the level he/she has entered for, the set of descriptors, presented in the teachers guide as a 20 cell grid (5 categories, 4 levels) make up an analytic scale of proficiency (Shohamy 1981: See Section 4.8). The RSA has also developed a series of examinations for foreign languages on a similar model, further enriched by the experience of the modern languages graded objectives and profiling movement (Royal Society of Arts 1989). The categories for Degrees of Skill this time are: Experiential Competence, Linguistic Competence, Rhetorical & Discourse Competence & Fluency, Socio-cultural Appropriacy, Strategies for coping with difficulties, Examination Consideration (in effect interlocutor support). Once again, although the philosophy is pass/fail mastery and the grid is once again presented with the categories vertically down the page, and the levels across from left

to right, the set of descriptors in effect makes up an analytic scale focusing on the aspects of proficiency selected by the developers.

There is also a second way in which examination levels can produce a scale of proficiency: when an examination institute chooses to present an existing suite of examinations as a scale. Cambridge have for a long time offered the First Certificate of English (1939) and the Certificate of Proficiency in English (1913), and in 1980 they added an examination called the Preliminary English Test based on Threshold Level. The RSA EFL examination series referred to above offered communicative alternatives to these exams during the early 1980s, and now that Cambridge have taken over the RSA communicative series and plugged the gap between First Certificate and Proficiency in both suites of exams, they are in a position to offer examinations in two styles at 4 levels. A new initial examination called "Key English Test" at Waystage gives a 5th level to the scale. This 5 level scale has since been adopted by ALTE (Association of Language Testers in Europe) which aims to establish a common framework for examinations in the European Community (ALTE 1993:1). The resultant series of levels cannot yet be regarded as a scale, since performances at the different levels cannot yet be related to one another (see discussion in rating scales above), but (a) Cambridge are in the process of calibrating their bank of Use of English items with the Rasch model and (b) work on the development of a series of descriptors for the levels is being undertaken.

1.3.3. Stages of Attainment:

A third origin of scales of language proficiency is the definition of stages of attainment as part of a framework of objectives, assessment and certification for an educational system or course of instruction. In this case the scale may take the form of either degrees of skill in performance outcomes and/or report a holistic characterisation of the type of language the person has at each level and the kinds of things they can do with it. Other elements may also be included. Such scales are a holistic overview of outcomes from graded levels, and they may be developed pragmatically in relation to:

(a) units of notional "seat time", e.g. 100 hours (original Eurocentre aim); school years (English National Curriculum).

(b) a series of specified exit points for different students for different languages for different purposes, as in the new Dutch framework. This attempts to define suitable objectives functionally to represent "distinct chunks of language competence (Van Els 1992:113).

(c) a series of levels considered to be critical to end-users of the education system (i.e. employers), as in the UK National Language Standards, or in the "Critical Levels Project" planned by LINGUA in co-operation with ALTE.

It is a feature of scales of this type (except perhaps the "critical levels") that they are process as well as product motivated. They can be a starting point for the generation of a coherent system of objectives, or the end point abstracted from such a system of objectives, or a synthesis arrived at from the two. Scales which describe stages of attainment thus tend to be have detailed content specifications in addition to the descriptors of degrees of skill in performance making up the scale. The first scale to have learning content specifications appears to have been the Stages of Attainment Scale developed by the English Language Teaching Development Unit, then the R & D arm of the ELT division of Oxford University Press (ELTDU 1976). The ELTDU scale claimed inspiration from Threshold Level, set up a series of 8 levels (of which the third and fourth were considered to reflect the Threshold Level content) and applied a similar form of task analysis as far up the scale as was felt to be feasible. The result of this analysis was a set of language specifications, which acted as a teacher guideline. ELTDU acted as consultants in creating the first version of the Eurocentre scale and language specifications in 1983, for which a similar approach was used.

In the same way that not all rating scales are scales of proficiency, however, not all schemes of stages of attainment are scales of proficiency either. The UK graded objectives schemes of the 1970s–1980s did not develop holistic descriptors of the target levels. However, now that such descriptors have been provided through the National Curriculum, largely by synthesising the content of the schemes (Page personal communication) those local schemes which survive are apparently adapting them for profiling and self assessment by focusing on the section of the continuum on the main

framework scale (a range of three or four levels) for the particular group of students concerned (Thorogood 1992:11).

1.3.4. Synthesising Approaches

In developing a transparent and coherent educational framework there is a certain tension between proficiency and achievement interpretations of attainment which the Swiss approach to the Language Portfolio mentioned in Chapter 1 attempts to resolve. Brindley outlines three different "levels" of achievement with which he proposes replacing the traditional proficiency/achievement distinction (Brindley 1989:10–18; 1991). In Brindley's account, (i) the view of people external to the program—employers, parents—learners as clients—is basically the "proficiency" view represented by scales of proficiency, and (ii) the view taken by teachers is more concerned with the enabling knowledge and skills which are focused on in the syllabus and is basically the "achievement" view represented by traditional teacher grading and mastery learning interpretations of criterion-referenced assessment. Brindley proposes a third view, which he suggests as a way to link the other two views in a perspective which he claims can be of interest to all partners. This third view refers to the "achievement of particular communicative objectives" (Brindley 1991:156) which can be assessed informally through self and teacher assessment using sets of descriptors presented in profiles, grids, checklists, and focus wheels (Brindley 1989: 47–120)

It is this third view of communicative achievement which the Swiss project hopes to promote with the Portfolio. If objectives are expressed in practical communicative terms they can provide a transparent and coherent metalanguage simple enough to be shared by end-users, parents and learners as well as teachers (Schneider and Richterich 1992). Checklists of very detailed Graded Objectives-style statements (e.g. I can introduce myself; I can say where I live; I can say my address in French, I can say how old I am, etc. & I can ask someone what their name is; I can ask someone where they live, I can ask someone how old they are etc.) can be summarised in a more holistic reporting statement (e.g. Can ask for and provide personal information) for a particular level on a subscale for Exchanging Information). In a similar way, entries on particular sub-scales at a particular level can then themselves be summarised into a holistic statement (e.g. Can exchange greetings

and interact in a simple way. Can ask and answer simple questions, initiate and respond to simple statements in areas of immediate need or on very familiar topics.) on a holistic reporting scale for Speaking—or Spoken Interaction. The Eurocentres Scale of Language Proficiency works in this kind of way with the global scale from which an extract was presented at the beginning of this chapter as the apex of the information pyramid (North 1993c).

Linking classroom interpretation of communicative achievement (bottom-up) to real-world demands for communicative proficiency defined at a series of levels (top down) can provide transparency and coherence in the overall system. Unfortunately though, existing systems rarely seem to manage to combine both aspects, as the following two examples suggest:

Both the ACTFL and ALL (Australian Language Levels) projects offer examples of cases where common frameworks have concentrated on one side or the other and thus failed to achieve a transparent, coherent system. The effect is that the US has a common framework, but one more relevant to teachers' qualifications than learners' stages of attainment, and that Australia has a dynamic set of stages of attainment and curriculum guidelines, but without a common framework of qualifications.

- ACTFL: The ACTFL Guidelines are an example of a top-down proficiency-oriented development in a political and in a linguistic sense,—adapting an existing holistic instrument on a government initiative to provide holistic descriptions of the learning product, and test tasks operationalising it. The background to ACTFL is to be found (a) in the Modern Language Association (MLA) Guidelines for Teacher Education Programs in Modern Foreign Languages (1966), (b) in the extension of the use of the FSI and adaptations of it to screen Peace Corps candidates in the late 60s and to assess high school modern language students (e.g. New Brunswick: Edwards 1985) and ESL students (e.g. Florida: Dade County 1978) in the late 70s, and (c) in the ETS "Common Yardstick Project" (Liskin-Gasparro 1984b). However, the immediate origin was the Presidential Commission on Foreign Languages and International Study in 1978, which recommended "a National Criteria and Assessment Program...to develop foreign language proficiency tests, and to report on, monitor and assess foreign language teaching in the US" ...and to "establish language proficiency achievement goals for the end of each year of

study, with special attention to speaking proficiency" (Liskin-Gasparro 1984b:29). Despite the attempt to "take the functional/notional (Council of Europe) emphasis on language in context for a particular purpose and apply it to the academic setting in the United States" (Liskin-Gasparro 1984b:35) the ACTFL approach is outcomes-oriented, basically confined to interview assessment, like the parent ILR system. The result is that in a context where university foreign language majors only achieve a level of 2 or 2+ /ACTFL Advanced Plus—roughly Cambridge First Certificate level—according to Carroll's survey (Carroll J.B 1967) or only 1+ /ACTFL Intermediate High—roughly a highish interpretation of Threshold Level according to Valette's more recent (1991) judgement—at the end of their university career, one could argue that the emphasis on assessing the end product rather than increasing and enriching the content is, as Barnwell and others have argued (Barnwell 1991), irrelevant to what is happening in schools in the United States.

- ALL: The Australian Language Levels Project provides an example of the reverse. Again the brief was to synthesise an input & achievement approach (from GLAFLL: Graded Levels of Achievement in Foreign Language Learning: the Scottish slightly more holistic version of the English graded objectives schemes) with an output, proficiency, calibration approach (from the ASLPR) in order to provide a national framework (Clark 1987:187). In this case, the proficiency, calibration, aspect of the brief was dropped and the project concentrated on providing graded levels of objectives, a communicative activities syllabus, and teacher guidelines on methodology—all the process aspects which the ACTFL approach lacks. But despite the brief, no scales of proficiency summarising the content and defining degrees of skill was provided as a framework of reference. The result is apparently an inability to relate the results in one state to those in another, and a new round of scale development, at any rate for ESL .

1.4. Purposes:

As the different origins outlined above suggest, scales are not all written for the same purpose. Alderson introduced a three-way functional classification of scales of language proficiency (Alderson 1991a:72–4):

(a) user-oriented, with the function of reporting: "information about typical or likely behaviours of candidates at any given level";

(b) assessor-oriented, with the function of guiding the rating process—typically expressed in terms of aspects of the quality of the performance expected;

(c) constructor-oriented, with the function of guiding the construction of tests at appropriate levels—typically expressed in terms of specific communication tasks the learner might be asked to perform in tests.

Alderson is here discussing scales purely in an testing context. When scales are used as an educational or training framework (e.g. as in the ELTDU and the Eurocentres case) rather than in conjunction with a test, then Alderson's category of constructor-oriented information could be expanded to also cover the "language specifications" attached to the scale, consisting of lists of tasks implied in the scale descriptors for each level, and language deemed necessary to perform the tasks at the level concerned. Such information can be used to inform the construction of syllabuses and materials and continuous assessment checklists as well as tests.

As Alderson points out, when the orientation of the scale does not fit the purpose it is actually used for, problems result. Pollitt criticises the inclusion in an assessor scale (ACTFL) of constructor-oriented task information rather than assessor-oriented definitions of the degree of quality in different aspects of performance (Pollitt 1991:88) and Fulcher suggests that doing so limits the generalisability of the test result to similar tasks and situations (Fulcher 1993:24).

Pollitt & Murray (1993) take Alderson's line of thought a significant stage further in pointing out that whereas many assessor-oriented systems define each aspect of performance for each level, creating a fully completed grid with levels as the vertical axis and aspects or qualities as the horizontal axis, in fact assessors appear to concentrate on different aspects at different levels. They propose a methodology to elicit what those salient features are. Detailed description of a particular aspect for a level at which it is not salient, whilst it may be useful to report a profile for diagnostic

purposes, is not assessor-oriented since, if anything, it complicates rather than facilitates the assessor's task—as Matthews (1990) complains. Pollitt & Murray therefore suggest the term diagnosis-oriented to describe scales which have these comprehensive descriptive grids. This suggestion is reinforced by research from the field of work evaluation which suggests that the addition of detailed descriptors for different performance dimensions leads to a primarily qualitative gain in improved feedback rather than a quantitative gain in the reliability of ratings.

To summarise, scales of language proficiency can thus be seen as having one or more of the following orientations:

- user-oriented

1. What the learner can do:

- constructor-oriented
- assessor-oriented

2. How well he/she performs:

- diagnosis-oriented

All four orientations can be considered relevant to a framework which seeks to provide a common defined point of reference for different educational contexts and perspectives. There will be occasions when only very simple, generalised statements are required for reporting results to non-specialist users (user-orientated). There will be other occasions when a detailed description of what a learner should be able to do for a particular purpose will be useful in order to identify priorities in the design of a learning module,—or in order to describe how the weighting of an existing model fits into a broader context (constructor-orientated). Learners and their teachers may find it helpful to be able to map or profile in relevant categories the progress being made towards a particular objective, and identify strengths, weaknesses and areas which for any number of reasons might represent a personal goal (diagnosis-orientated). Finally, achievement may be assessed for certification in relation to specific standards defined in terms of levels of the framework (assessor-oriented).

The development of a scale of proficiency which can cater for the different purposes and perspectives outlined above is a relatively complex operation. In order to help in the identification of priorities for the development of syllabus, activity and/or test design—i.e. in order to help in the development of content specifications, or in the establishment of the relationship of sets of such content specifications to each other (constructor-oriented)—the scale will need to offer descriptors for the kinds of communicative activities likely to be relevant. In other words, in terms of Bachman's (1990:303–330) classification, a common framework scale needs a "real life" dimension. Such coverage should ideally be related to theory, if this is possible given the state of development of relevant theory. It could also be argued that such a scale should also be based upon a needs analysis, or at least on a needs analysis methodology in the way in which Carroll (1979) used Munby's (1978) needs analysis model to arrive at the original specifications for the English Language Testing Service—for learners wishing to attend a British university (See Spolsky 1986:155 in this respect). The Eurocentres scale, which operates as a framework for the Eurocentres schools teaching languages in the countries in which they are spoken, is based upon such a needs analysis even though the orientation of the scale is general rather than specific purpose (Johnson and Scott 1984). However, firstly the development and implementation of such a needs analysis is full of theoretical and operational pitfalls (See Criper and Davies 1986 in relation to ELTS). Secondly (I)ELTS and the Eurocentres scale are each used for particular purposes in particular educational sectors, whereas a common framework scale needs by definition to be as relevant as possible to all sectors making a needs analysis almost contradictory (how can you analyse all possible needs?). Thirdly, needs analyses of this sort have in any case been criticised for creating a static view of both the goals and processes of language learning (see e.g. Hawkey 1980; Maley 1980; Ladousse 1982; Robinson 1983; Davies 1990:135). Finally, the major potential use of a common framework scale is to serve as an instrument providing a neutral profile of proficiency—i.e. one that is comprehensive and related to theory — which can then be used to profile the needs of particular groups (See London Chamber of Commerce in Trim 1978 for an example of how this can be done diagrammatically). For these reasons, the needs analysis approach has not been applied in this project, though successive groups of teachers have been asked in a systematic

fashion about the relevance of particular descriptors to their learners' needs.

In order to help in the identification of priorities in relation to qualitative aspects of language use needed for adequate participation in particular communicative activities in particular domains, and in order to be able conduct a language audit (registering unique strengths and deficits) on these qualitative aspects as well as on the activities themselves, (diagnosis-oriented)—the scale will need to offer descriptors for aspects of communicative language proficiency and strategic competence. In other words, in terms of Bachman's (1990:303–330) classification, a common framework scale needs an "interactional/ability" dimension. As pointed out by North (1993d:7); McNamara (1995:159–165) and Brindley (forthcoming: 21) this entails a certain tension between incomplete theoretical models on the one hand and operational models developed by practitioners on the other hand. An attempt has been made in this project to relate the categories of description employed on the one hand to theory, and on the other hand, since as Skehan (1995a) points out theory has considerable difficulty accounting for the "ability for use" which is of primary interest to less specialised users to such operational models.

1.5. Types:

The distinctions made by Bachman (1990), Alderson (1991) and Pollitt & Murray (1993) are not the only ways in which scales of language proficiency have been classified. In situating the Australian Second Language Proficiency Ratings in the context of other approaches for a teachers' conference, Ingram and Wylie (1989) present the concept of scales of language proficiency as a series of dichotomies. As in many classification systems, the different dichotomies overlap somewhat, but give a rough-and-ready overview of the range of different approaches. The following table reorders Ingram and Wylie's categories slightly, and adds a short explanation and example.

Whole range of proficiency

Scales which go from zero to what is normally described as native-speaker or "near-native" proficiency.

Examples: FSI (Foreign Service Institute); ILR (Interagency Language Roundtable); ESU (English Speaking Union) Framework Project (Carroll and West 1989).

Partial, relevant range

Scales which concentrate just on the relevant section of the proficiency spectrum, from zero to the maximum relevant level.

Examples: English National Curriculum, which covers the range of proficiency of English school-children in 10 levels; Eurocentres, with 10 levels up to approximately the level of Cambridge Proficiency.

General proficiency

Scales aimed at learners without specific narrowly definable purposes in learning the language.

Examples: ACTFL; Eurocentres (both of which use a 4 skill model).

Language for specific purposes

Scales aimed at specific usually work-related activities presumably identified with a needs analysis.

Examples: ELTDU (English Language Development Unit) and IBM France 1974 (included in Trim 1978); ALTE (Association of Language Testers in Europe).2

Four skill

Scales which offer sub-scales for Listening, Reading, Speaking (or Interaction) and Writing—with or without a Global Scale.

Example: all the above except ELTDU and IBM.

Overall

Scales which offer just one scale for global language proficiency.

Example: Finnish Foreign Language Diploma for Professional Purposes.

Serial

Scales which operate a continuum model: if someone is a very good Level 5, one checks is they are Level 6.

Examples: ILR/ACTFL, ASLPR & Eurocentres.

Threshold

Scales placing a series of pass/fail thresholds in relation to one another.

Examples: RSA/Cambridge Certificates in Communicative Skills in English; RSA Modern Languages; ALTE.

Global, holistic

Scales requiring a synthetic judgement, balancing all factors: which level is most applicable.

Examples: ASLPR, Eurocentres.

Absolute/noncompensatory

Scales on which the candidate has to pass all of a number of specific points or criteria: fail one and you fail the level.

Examples: ILR/ACTFL; RSA/Cambridge.

Total behaviour

Scales which are "quantitative and qualitative" (Ingram & Wylie) i.e. they describe both the kind of task involved and the expected degree of skill in the performance

Example: FSI; Eurocentres ASLPR (to some extent).

Tasks only

"Can do" scales which restrict themselves to saying what the learner can (and cannot) do, but which do not mention quality of performance.

Examples: ELTDU; ALTE.1

Looking at scales of language proficiency from a different perspective in a detailed survey, North (1993a) presents scales in five groups:

i. Brief, holistic scales of overall competence in spoken interaction

ii. User scales reporting competence in different contexts of use

iii. Detailed, holistic rating scales

iv. Analytic rating scales

v. Frameworks of syllabus content and assessment criteria for stages of attainment

i. Brief Holistic Description: The "mother scale" the FSI could be regarded as belonging this category because the descriptors are short and user-friendly. In the survey the FSI is placed in the third group as the head of one of the two main "families" of scales in this group, the FSI family. The other scales in the first category share the FSI quality of being short, holistic user-friendly statements for each level; several date from the late seventies, but others are more recent like the Ontario ESL Oral Interaction Assessment Bands, (Canale 1984; Wesche 1992) or the Finnish Nine Level Scale of Language Proficiency (Luoma 1993).

ii. Different Contexts of Use: The second category, with a functional rather than skill orientation, was pioneered in the 1970s by three LSP (Language for Specific Purposes) projects: The ELTDU Stages of Attainment Scale, originally developed for the Swedish company SKF (Aktiebolaget Svenska Kullagerfabriken) (ELTDU 1976); the Canadian Language Selection Standard: Determining the Linguistic Profile of Bilingual Positions (Public Service Commission of Canada 1977) and the scale developed by IBM France (IBM 1974: appendix in Trim 1978). Two examples for general language stem from the Council of Europe project, an example for Social Skills (Trim 1978) and a proposal for an alternative to the four skills (North 1992a).

iii. Holistic Rating Scales: The detailed holistic rating scales in the third category include the FSI and Carroll/IELTS "families",

which account for 90% of the literature on scales of language proficiency. The FSI family encompasses the FSI, ILR, ASLPR and ACTFL which all share the same levels, much of the same wording and the same philosophy—except as noted above that ASLPR is a holistic judgement whereas ILR/ACTFL is "noncompensatory". The second main "family" is the B.J Carroll and ELTS (English Language Testing Service) family (Carroll 1978/81;1980). Another detailed holistic scale is one highlighted by Van Ek (Elviri et al, in Van Ek 1986) because it systematically incorporates information related to underlying aspects of competence.

iv. Analytic Rating Scales: These can be seen as deriving from the "factors" considered during the FSI oral interview, and scored on a checklist (not corresponding to the 0–5 FSI levels) at the end of the interview to serve as a point of reference in case of disagreement between the two examiners. Shohamy (1981) introduced the distinction holistic : analytic, and like other pre-communicative examples, her scale stays close to the FSI "factors" Grammar, Vocabulary, Fluency, Pronunciation (accent in FSI). A fifth FSI factor "comprehension" used to assess listening in the interview, appears early adaptations of FSI to school contexts (e.g. Dade County 1978). Analytic rating scales appearing in the communicative era show considerably less consensus on the "factors" involved. Some take a performance orientation (e.g. Carroll 1980: 137ff), others focus on aspects of communicative competence suggested by Canale and Swain (1980, 1981) (e.g. Gotheborgs Universitet undated–late 80s). Some, including both those examples, expect the rater to juggle with an astonishing number of categories (10 in Carroll's case; 2 holistic plus 9 analytic in Gothenburg's case) recalling Matthews (1990) complaints about feasibility and Pollitt and Murray's (1993) distinction between assessor-oriented and diagnosis-oriented scales (see Section 2.4.).

v. Educational Framework Scales: Scales in the fifth category stem from stages of attainment in educational systems and can, like the Eurocentres scale and the British National Language Standards (Languages Lead Body 1992) encompass all the types so far mentioned. In addition, they may contain detailed content as well as outcome specifications with links made between the holistic statements appearing in the scales themselves, and guidelines consisting of lists of tasks, functions and sometimes structures and vocabulary considered appropriate as content for each level (e.g.

Eurocentres and the RSA Modern Languages). Such definition in content specifications tends to be confined to lower levels in deference to the fact that global or holistic; functional, and structural views of proficiency are complementary and cannot be mapped in one-to-one relationships (Spolsky 1989:79).

1.6. Metaphors for Scales:

In an attempt to facilitate the introduction of what for their audience of teachers may have been a novel topic, Ingram and Wylie (1989:2) use the metaphor of the scales in a shop like a green grocers: "A proficiency scale selects certain graded criteria against which to "weigh" learners ability to use the language, it selects criteria that can be graduated between two points so that the learner's ability to use the language or some relevant aspect of their language can be matched against the criteria themselves. Generally a number of intermediate points are selected and criteria assigned to them so that a "set or series" of graduated steps is provided between the two end points." This appears rather an unusual interpretation of the word "scale" since it focuses on the way a set of weights and measures in a green-grocers swings between the two points fixed by the weights and the goods until a balance is reached, rather than upon the measurement scale marked in equal intervals called grams, ounces or whatever. Whilst this very much reflects the particular procedure adopted in the oral proficiency interview associated with the ASLPR, ILR and ACTFL with the "level probe" to the point of "linguistic breakdown", followed by the swing back to a more comfortable conversation (See for example Ingram 1985 for a detailed description), it represents the process as a series mastery/non-mastery decisions in relation to the thresholds between bands rather on placing the subject on the continuum marked out by ruler provided by a measurement scale. This draws attention to the point made at the beginning of Chapter 1 (and in North 1993d) that the extent to which scales of language proficiency have been discussed without reference to measurement scales or scaling theory is really quite surprising. This issue is discussed in more detail in Chapter 5., the point about mastery/non mastery decisions in contrast to situating the subject on a continuum in criterion-referenced assessment being taken up in detail in Section 5.1.

A second metaphor has been used by Schärer (1992) at a second conference, the Rüschlikon Symposium from which this project

derives. This metaphor involves the analogy of a mapping scale (e.g. as on UK Ordnance Survey) or modelling scale (e.g. as with model railways and other scale models). A map is used for orientation. You find your way around a map by using squares established with a horizontal and vertical axis. The scale here is (a) the degree of reduction or expansion—of specificity—also used in architectural models, scale models or cars, aircraft etc. etc. , and (b) the related unit of measurement. It is frequently the case that maps (and even models) use two or more different but related units of measurement. For example when driving to Zürich on the motorway you consult a map with a very large scale; as you approach the town, you might flip the map over to a more detailed representation of the conurbation, and when you go walking in the town centre, you use the town plan printed in one corner which uses a far more detailed scale and may even show you individual buildings.

This "map" metaphor was used to describe the second, informal section of the proposed Language Portfolio in which the learner can keep a personal record of learning experiences. In an educational context a grid of diagnosis-oriented descriptors (Pollitt & Murray 1993) can have a valuable formative function in that "it acts as a map for the students. They can see where they are and where they are headed. They like to see a map of objectives , but this is probably only a real virtue at the formative stage. It probably has no virtue at all (compared to a checklist presentation) at the summative stage. " (Stratton 1986:118). When such grids are used with a vertical axis representing a scale made up of "waystages" like those of the Scottish GLAFFL (Graded Levels of Achievement for Foreign Languages) project (Clark 1987:143) there is, even from a measurement point of view, no assumption that anyone should in fact follow the notional progression of the psychometric scale used to express the vertical dimension of a profile grid. "Since the scale can be understood as an ascent from the bottom, descent from the top, or digression from the centre, the levels specify a conceptual hierarchy, but do not require that a path must be followed" (Linacre 1991:155). Development can be lateral as well as vertical, language loss can be profiled as well as language gain.

Grids which function as orientation tools in this way focus on what is deemed useful to the degree of specificity deemed optimal, with a degree of user choice. Douglas (1988:257) criticised the holistic ACTFL scale for its inability to register sideways movement and

contrasts it to what he calls the European perspective encapsulated by Trim's (1984:20) statement: "A learning biography consists not of a straight line progress from elementary to intermediate to advanced but an accumulation of life-related learning experiences." This implies a more modular approach: a grid of categories on which to chart (map) sideways and upwards progress.

The concept of "mapping" areas mastered in relation to a metalanguage of descriptors for different types of relevant categories is consistent with the purposes to which a common framework scale may be put and is consistent with measurement theory (Linacre 1991, cited above) provided that certain concerns about the degree of unidimensionality or multidimensionality are taken into consideration.

1.7. Descriptions of Behaviour, Behavioural Objectives & Utilitarianism

Such mapping charts learning outcomes. All scales of language proficiency are specifications of outcomes, generally expressed in terms of tasks the learner can perform (constructor/user-oriented; "real life") and/or the degrees of skill in various aspects of performance (assessor/diagnosis oriented; "interactional/ability"). Hence scales of language proficiency are "behavioural"—an adjective many teachers feel wary about—in the sense that they describe behaviour. Because of the concentration on outcomes, because of the use of behavioural definitions, scales of language proficiency have sometimes been interpreted as representing a utilitarian, behavioural, or even behaviourist perspective. This conclusion is somewhat exaggerated: scales focus on behaviour because they tend to take a functional view of proficiency, describing what people can do. A functional orientation does not, however, imply an exclusive focus on performance testing and work sample collection but is rather perhaps a continued reaction against intellectualising language learning expressed as "teach the language not about the language" (Halliday, MacIntosh and Strevens 1965:254, cited by Stern 1992:80). "Language learning has two sides to it: knowing and doing (competence and performance)...different approaches to language teaching have tended to emphasise one rather than, and often at the expense of the other" (Widdowson 1990:157) and it is possible that the "proficiency movement" like the "communicative movement" may

have led in some contexts to the unwarranted conclusion that because we are talking about "X" we can now forget all about "Y". In Britain, communicative proficiency is in fact not necessarily tested exclusively through "real life" performance testing (Bachman 1990a:303). ELTDU and Eurocentres have a "system knowledge" test as part of the procedure which place peoples on their scales; IELTS developed one and only dropped it since it appeared not to add to the information available by aggregating the results from other tests.

The fact that statements on proficiency scales are written in terms of holistic behaviour, and thus are in this sense objectives expressed in behavioural terms, has in fact nothing whatsoever to do with "behavioural objectives" in the sense of the behavioural objectives movement of the 1960s and 1970s which define micro pedagogic objectives which as a general rule bear little relationship to real life performance. The classic example of a behavioural objective applied to language learning is cited by Stern (1992:67):

> *"To demonstrate knowledge of twenty out of fifty vocabulary words ...write out and spell correctly the word that corresponds to each of the twenty definitions given on a twenty-minute classroom test. At least thirteen of the twenty items must be entirely correct in order to pass.*
>
> Valette and Disick (1972:17)

The fact that Mager's influential work marrying the idea of a performance standard to the 1930s behaviourist idea of a behavioural objective (Mager 1962) was published virtually simultaneously with the seminal work on criterion-referenced testing (Glaser 1963) led to a fusion of the three approaches in the US into so-called "mastery learning" and the definition of "minimum competence standards" with which criterion-referenced testing has been over-identified in the US, some would say disastrously so (Jaeger 1989:488) despite the fact that "this language of performance standards is pseudo quantification, a meaningless application of numbers to a question not prepared for quantitative analysis" (Glass 1978:238) and "there were in Glaser's early writings few intimations that criterion-referenced tests could be used to establish cut-off scores between competence and non-competence or that such distinctions as pass/fail and mastery/non-

mastery make psychological sense" (Glass 1978:240). Glass concludes "The evolution of the meaning of "criterion" in criterion-referenced testing" is, in fact, a case study in confusion and corruption of meaning" (Glass 1978:242).

A second way in which scales of language proficiency have been misinterpreted is in identifying them with professional training and social engineering undertaken at the expense of personal development. Scales of language proficiency, as expressions of functional goals, do seem to belong squarely in the "reconstructionist" (i.e. social reforming) curriculum perspective in Skilbeck's (1982) classification: classical humanism; reconstructionism; progressivism) cited by Clark (1985:344; 1987:14ff), but this does not necessarily imply a utilitarian "training" view of language goals. In any case, the distinction between training and education can be overdone: "Training is akin to following a tightly fenced path, in order to reach a predetermined goal at the end of it. Education is to wander freely in the fields to left and right of this path—preferably with a map...As most training involves some unplanned learning (educational effects) and most education involves some planned goal-orientated teaching, the value of these two terms as discriminators is somewhat dubious" (Romiszowski 1981:3, my italics). If the learner is going to develop some autonomy, he/she is going to need a map of the kind discussed in Section 2.6. above which statements in profile grids or circles, scales and checklists can provide. Training in map reading—in how to use a metalanguage to organise learning and recognise stregths and weaknesses—is a prerequisite for self direction (Oscarson 1978;1988;1989; Dickinson 1987). As Kohonen (1989, 1992) has also pointed out, product and process approaches to evaluation are necessarily complementary and there is no particular reason why the provision of holistic definitions of behaviour at different levels should preclude progressive approaches promoting autonomy and humanistic methodologies. They have in fact been used to promote them in a Council of Europe context (Oscarson 1984) and in the British Profiling movement (Thorogood 1992:2–9). One should not overlook the distinction in needs analysis between target or terminal objectives implied by scale definitions on the one hand and ongoing classroom negotiation, using instruments which may be related to the scales on the other hand. This dual focus, this so-called two step approach (Hawkey 1980:91; Clark 1987:37), has been a fundamental feature of the Council of Europe approach

(Richterich 1983; Oscarson 1978; 1984;1988), even if this fact was sometimes overlooked in British literature in the early 1980s reacting against the "static study of inter-role relations" (Davies 1990:135) represented by Munby's (1978) restrictive ESP model.

However, it does remain true that a scale of proficiency can only include what people have been able to define in words and assign on some basis to different levels of proficiency. It may also be true that "in school language teaching it has been the case that the communicative objectives that get specified are often highly transactional in nature (buying, getting tickets etc.) and that the more expressive and creative functions of language, which are more difficult to set out in terms of behavioural objectives, get left out" (Clark 1985:347). However there is no particular reason why this should in fact be the case, as more recent attempts to provide specifications for foreign language learning suggest (e.g. Hébert 1990). Descriptions of the creative and expressive side of at least writing are often included in proficiency scales for mother tongue language learning (e.g. Quellmalz 1982a, 1982b) and, again, there is no particular reason why they should not be included in scales for foreign language learning too, and why a comprehensive common framework should not accommodate them.

The identification of scales of language proficiency with behavioural objectives and utilitarianism can therefore be argued to be misconceived. That certain scales may concentrate on highly transactional tasks doesn't mean that all scales must do so. A common framework is likely to be exploited in a number of different ways from a number of different perspectives and should therefore be as comprehensive as possible, incorporating an "interactional/ability" perspective as well as a "real-life" perspective. This is an ambitious task in the development of any scale, and when it is undertaken in regard to a scale which is intended to be referred to in a number of contexts, the complexity of the problem naturally increases.

2. Difficulties With Common Framework Scales

2.0 Introduction

One of the major problems in relation to the development of a common framework scale is that it should be as comprehensive as

possible and it should be possible for different users to relate their own scales and sets of levels to it. This makes it difficult for any one person or group of people to write it, and since the scale needs to have properties which are generalisable, it needs to be related theory. Most scales of language proficiency appear to have been produced pragmatically, by appeal to intuition and those scales which already exist with little consideration of theory (Brindley 1991:6–8). Whilst this approach may be appropriate in the development of an in-house system for a specific context with a familiar population of learners and assessors, it has been criticised in relation to the development of national framework scales (e.g. Skehan 1984; Fulcher 1987, 1993 in relation to the British ELTS; Brindley 1986, 1991, Pienemann and Johnson 1987 in relation to the Australian ASLPR; Bachman and Savignon 1986, Lantolf and Frawley 1985, 1988, 1992; Spolsky 1986, 1989, 1993 in relation to the American ACTFL).

The problem can reformulated in the following way. A scale of proficiency can be said to have two axes, a horizontal axis—categories, and a vertical axis—levels or bands. In other words there is a description issue,—that the categories employed are related to a model of competence—and there is a measurement issue—that since everyone will treat the scale as if it is linear, it should be related to a model of measurement. It is sometimes artificial to draw a line between the two sides of the problem, description and measurement, but the distinction is used for convenience in the organisation of this thesis. Some of the central requirements for a common framework scale are the following:

Description Issues

- A common framework scale needs to be context-free in order to accommodate generalisable results from different specific contexts, yet at the same time the descriptors need to be context-relevant, relatable or translatable into each and every relevant context—and appropriate for the function they are used for in that context. This means that the categories of description to describe what learners can do in different context of use must be relatable to the target contexts of use of the different groups of learners within the overall target population.

- The description also needs to be based on theories of language competence, although the available theory and research is inadequate to provide a basis for it. Whilst relating to theory, it must also be relevant to the contexts of the learning population concerned, and it must remain user-friendly—accessible to practitioners—and should encourage rather than discourage them to think further about what competence means in their context.

Measurement Issues

- A common framework scale needs to have scale values which are based on a theory of measurement in order to avoid systematising random error in the system itself through adopting unfounded "rules of thumb" of either existing scales, the authors, or groups of practitioners consulted.

- The number of levels adopted should be adequate to show progression in different sectors, but, in any particular context, should not exceed the number of levels people are capable of making reasonably consistent distinctions between. This may mean adopting different sizes of scale step for different dimensions, or a two-tier approach between broader (common) and narrower (local) levels.

2.1 Description Issues

2.1.1 Context-free : Context-relevant:

A common framework entails providing a set of descriptors which are context free—and yet to be effective those descriptors should be capable of being related to any relevant context, and of being interpreted reasonably consistently across those contexts by different groups of users (Nuttall & Goldstein 1986). There are two issues here:

Firstly it can be argued that the concept of "proficiency" as it is described in rating scales such as the ACTFL or ASLPR is context dependent. In other words, if proficiency is defined in terms of people's ability to use language for particular communicative purposes, as is now the case, then the criteria for "proficiency" which would be applied in the case of adult immigrants in Australia, for example, would be different to those used for a group of graduate students in the United States (Brindley 1991: 154–5).

Spolsky also takes up this argument: "A functional set of goals exists in a social context." "Where this is consistent and common as in the Foreign Service, or in the Council of Europe notion of the Threshold Level for tourists and occasional visitors, it is not unreasonable to develop a scale that proceeds through the skills."..."If it cannot be based on a single social goal, a single set of guidelines, a single scale could only be justified if there were evidence of an empirically provable necessary learning order, and we have clearly had difficulty in showing this to be so even for structural items" (Spolsky 1986:154; 1989:65).

This argument would appear to confine scales of language proficiency to LSP—counting being a tourist as a specific purpose. Yet, firstly, the eighties saw a widespread disillusionment with the "specific" form of language for specific purposes (e.g. Munby 1978) as it was discovered that teaching more generalisable functional skills was more practical (e.g. see Ingram and Clapham 1988). Secondly, the Threshold specifications have been adapted successfully for other more specific contexts (e.g. immigrants), but also for other more general contexts (schoolchildren, learners on stays abroad) by curriculum developers, course book writers, examination boards. Must a functional definition of one or more stages of attainment be identified exclusively with LSP? This doesn't detract from the argument that scales need to be designed for the function they will be used for and that it is dangerous to lift scales defined to function in one context to describe a certain type of learner for a certain type of user, tinker with them and then use them in another context to describe a different type of learner to another kind of user. This ignores the domain specificity of the original scale with regards to raters and ratees—which is what Spolsky argues ACTFL have done in adapting the FSI (Spolsky 1986:150; 1993:208).

Spolsky's and Brindley's point could be reinterpreted as being that a communality of functional goals should be demonstrated, and not taken on trust. In this regard, the first step in the development of the Eurocentre Scale of Language Proficiency was a survey of perceived needs with a 30% representative sample of Eurocentres UK students.

The second aspect of this problem relates to the interpretation of the same descriptor by users in different sectors or regions. It would

be perfectly feasible to have functional goals which were in fact common to the contexts in question, and still have a scale which failed to operate as a meta-system because each group interpreted the same wording in a different way. Trim has drawn attention to the problem of descriptors which are "capable of an indefinite number of often contradictory interpretations, and so they can easily gain an apparent acceptance" (Trim 1978:56). Vagueness and/or norm-referenced relational description (a common feature criticised in scales of language proficiency e.g. Skehan 1984:217) can be expected to be misinterpreted in terms of the norms of the sector/examination concerned—thus losing comparability between sectors. It is in practice very difficult to avoid slipping into this kind of description, but by refining the descriptors with groups of teacher. an attempt has been made to develop descriptors which, whilst being generalisable, try to offer a transparent precision which provides points of reference for criterion-referenced assessment.

Even after one has (hopefully) avoided the pitfall of vagueness, there remains the problem that descriptors relating to particular types of tasks may be interpreted in a systematically different way in different sectors since they may be significant functional goals in the one context—and therefore practised, expected in learner performances and hence "easier", whereas they are not seen as so central to another context, are not focused on and hence are considered "more difficult"—Spolsky 's point cited above. However, this phenomenon, technically known as "differential item functioning" is routinely investigated in relation to test items when using item response theory (IRT) scaling methodology. Because IRT operates with individual items, by treating different descriptors as items and analysing teacher ratings with the Rasch Rating Scale Model (Wright and Masters 1982) it is possible to investigate how people in different contexts relate to the same descriptor—in other words it is possible to determine how context free it is. The many-faceted version (Linacre 1989, Linacre et al 1992) of the Rating Scale Model enables one to take account of and adjust for the subjectivity of judgements themselves and to investigate systematic variation in the interpretation of descriptors by demographically defined "facets" like educational sector and linguistic region—and hence to evaluate to what extent the framework of description offered is in fact common to the different contexts involved.

2.1.2. *Theory-based—User-friendly*

There are those who consider that the development of a common framework should not be attempted because research has not provided an adequate empirically validated description of the complexity of language proficiency (Lantolf and Frawley 1985, 1988, 1992). Spolsky has voiced a similar concern (Spolsky 1993:208).

In discussing the shortcomings and limitations of scales of proficiency, there is an important distinction which should be made between a theoretical model to describe the nature of foreign language proficiency, and an operational model which people can actually use. An operational model is always simpler than a theoretical model, and whilst it may relate to theoretical models, it may reinterpret elements to make them more accessible in a particular context. Even theoretical models do not describe reality, they "make ideas about experience explicit. They specify how experience might be simplified so that it can be remembered and managed" (Wright and Masters 1982:60) "in order to represent the crucial features of a complex situation, and should not be expected to be a true reflection of reality" (Choppin 1981:4).

In this sense, then, the criticism by Lantolf and Frawley (1985:341) that scales of proficiency (in this case the ACTFL Guidelines) model reality rather than mirroring it, that they have "constructed a reality" and are "prescriptions of a theorist deciding what speakers ought to do" is simply misguided. All models model reality: that is why they are called models and "we cannot wait for the emergence of empirically validated models of proficiency in order to build up criteria for assessing learners' second language performance" (Brindley 1989:56). As Hulstijn says: "it should be obvious that syllabus writers, teachers and testers cannot wait for full-fledged theories of language proficiency to emerge from research laboratories. In the absence of theories, they have to work with taxonomies which seem to make sense even if they cannot be fully supported by a theoretical description" (Hulstijn 1985:277). These arguments appear even more cogent if one takes the view that there will probably never be a fully generalisable empirically validated description of language proficiency.

Lantolf and Frawley criticise the ACTFL Guidelines because they are finely honed committee-produced "lovely symmetrical"

descriptors (1992:35). They consider that the descriptors have no validity because they are "groundless, made up—arbitrarily" (1992:35). However, the fact that a particular standard may be found to have decisions which can be criticised, the fact that a standard is "arbitrarily" set is not in itself an argument against it since all standards, all criteria are "arbitrary" value judgements whether they are fire standards, health standards or environmental standards (Popham 1978, Hambleton 1978, Cronbach 1961 cited in Davies 1988).

> *"Unable to avoid reliance on human judgement as the chief ingredient in standard-setting, some individuals have thrown up their hands in dismay and cast aside all efforts to set performance standards as "arbitrary" and hence unacceptable.*
>
> *But Webster's dictionary offers us two definitions of arbitrary. The first of these is positive, describing arbitrary as an adjective reflecting choice or discretion, that is "determinable by a judge or tribunal". The second definition, pejorative in nature, describes arbitrary as an adjective denoting capriciousness, that is: "selected at random and without reason". In my estimate, when people start knocking the standard-setting game as arbitrary, they are clearly employing Webster's second, negatively loaded definition.*
>
> *But the first definition is more accurately reflective of serious standard-setting efforts. They represent genuine attempts to do a good job in deciding what kinds of standards we ought to employ. That they are judgmental is inescapable. But to malign all judgmental operations as capricious is absurd."*
>
> Popham 1978:168, cited in Hambleton (1978:102)

The arbitrariness can be limited by (a) taking account of theory and research; (b) taking steps to ensure that decisions taken are based on a wide consensus in the relevant context, (c) by taking account of the subjectivity in individual and group judgements (through the many-faceted Rasch model: Linacre 1989), and (d) by stating clearly what the limits to empirically established validity are, since, rather than saying that a test, or scale, is valid, one should specify what it is valid for (Henning 1990:379). Nevertheless, in the absence of the

ultimate description of language ability, arbitrariness (in the positive sense) will remain.

Apart from the question of the inevitable incompleteness of any descriptions of proficiency in a common framework, there is the question of their accessibility to those people who will use them. "Concepts like communicative competence, sociolinguistic competence etc. are constructs. In other words, they are creations of applied linguists which, it is claimed, have some theoretical justification" (Skehan 1984 p 209). However, even if such concepts of underlying competence can be theoretically justified and defined, they are difficult to operationalise and observe which is one reason why most rating systems focus on simpler more observable aspects of performance like range, accuracy and fluency. Not only that, but there are indications that the particular model of competence used to rate performance on tasks may in any case not be too important in relation to the differences in performances caused by the different requirements of the tasks themselves (Pollitt and Hutchinson 1987:90). This is probably just as well since a cursory glance at a collection of scales of proficiency (e.g. in North 1993a) shows that rating categories vary tremendously; there are a myriad of factors, one can only work with a few, and so people group them in different ways in order to emphasise aspects they consider to be particularly important in the context concerned. Institutions develop their own criteria and train raters to use them, developing "schools" in the process. Some schools seem to think they have an exclusive definition of proficiency: most, however, recognise that experts take many routes to the same goal (Einhorn 1974).

Since rating categories are a metalanguage to talk about competence, and since (a) this is a very valuable experience for teacher development and (b) it is very difficult to change the way people think and form prototypes, there is an argument that the categories used should have relevance for the people who are expected to use them, should be presented in comprehensible, practical language which avoids the jargon of applied scientists and should preferably be developed empirically with representative informants. Developing assessment scales with the kind of people who were going to use them was the approach pioneered by Smith and Kendall (1963) who developed the first form of what are called generically behaviourally-based rating scales. They were reacting against a practice in which abstract categories (traits) determined

by psychologists on the basis of intuition or factor analysis were parachuted into hospitals to be used by head nurses in rating their juniors. The problem then as now is that sophisticated but opaque categories and/or complex theoretical jargon can be expected to be ignored in favour of the norms and rules of thumb of the sector/examination or person concerned. All evaluation of behaviour is heavily influenced by a tendency to match small amounts of vivid concrete information to preconceptions and prototypes which form implicit standards (Murphy et al 1982:563; Parks 1985:181; Murphy and Cleveland 1991:127, 150ff). If descriptive categories are too numerous or too complex, raters tend to fall back on their own prototypes (Matthews 1990a).

There is therefore an important pitfall to avoid in developing a descriptive metalanguage. From a practical point of view, there is not a lot of use in developing a theoretically sound but operationally very difficult set of categories since such an approach would be very likely to be dismissed by teachers as "theoretical, academic or airy-fairy" (Davies 1985:8). In such a case the teachers, as Matthews says, would then tend to retain simplistic prototypes which are of course based on outdated theory, since as McNamara (1995:164) points out "even practical approaches which try to eschew theory imply a theoretical position. This is often found in the criteria for assessment, which embody an implicit view of the construct"[2]. What is therefore required in order to develop workable descriptors for a set of categories that is informed by theory would seem to be a forum for dialogue between the practitioners and the theoretical categories. Teachers may not like airy-fairy ivory tower thinking, but they very much do like concepts which are new to them which ring true and which they see as relevant to the improvement of the quality of their learners' performance. One way to do this is the method used by Smith and Kendall (1963) in successive workshops with nurses, and that approach has been adapted in this project as described in Chapter 7.

2.2. Measurement Issues

2.2.1. Avoiding Systematising Error:

[2]In another context, the late Maynard Keynes is reported to have commented, in relation to the contempt most "men of action" appear to have for economic theory, that a businessman who so describes himself generally turns out to be the slave of some long dead economist.

Systematising the random error of method effect: A major criticism of the ACTFL system is that there is a considerable circularity of argument: "proficiency" is defined as what is tested in the "oral proficiency interview" which is defined as the operationalisation of the guidelines, which define proficiency—confusing the trait with the method: (Bachman and Savignon 1986:384, Bachman 1987:33; 1988). This is a problem with all subjective assessment systems which cannot make an adjustment for task difficulty and rater severity. It is particularly a problem with interviews which are a ritualised unequal encounter in which the interviewer is defending counsel, jury and judge all at the same time, in which the dominated partner has a restricted range of roles (Raffaldini 1988, Kramsch 1986, Van Lier 1989, North 1993b).

Systematising the random error of inappropriate "rules of thumb". Everybody has their prejudices: personal criteria which make short cuts in whatever the official system is; the question which "sorts people out"; the rule of thumb which says "I find that people who can do this are intermediate". The reason for having descriptions attached to levels is an attempt to make the criteria applied explicit, shared, and consistent. If a hierarchy amongst the descriptors describing the same dimension is not established on the basis of theory, experience and empirical item analysis, (Murphy and Constans 1987; Murphy and Pardaffy 1989), if a simplistic assumption is used to fix a key threshold or cut-off between two levels, (Landy and Farr 1983), this systematises the very kind of error one is trying to avoid, producing possibly consistent, but invalid measurement.

Using specific linguistic forms as a way of discriminating between levels as do the ACTFL Guidelines (e.g. consistent use of the past = advanced, as in ACTFL) rather than making a holistic judgement about range, accuracy , fluency etc. (as in FSI) (Savignon 1985:1003–4) is an example of systematising measurement error, since it posits a simple, causal relationship between proficiency level and correct production of a particular form. This ignores the fact that although SLA (Second Language Acquisition) research has shown underlying systematicity in developing interlanguage, it shows itself in the emergence of particular forms rather than accurate mastery of them—which may be subject to many influences. Meisel, Pienemann and Johnson (1987) following Clahsen (1985) posit two dimensions in their Multidimensional Model: (i) the stage of acquisition (yielding

fixed sequencing) and (ii) the orientation of the learner (including demographic characteristics) As Ellis (1989a:310) points out there is considerable variation "some learners (dubbed error-avoiders) seek to master a rule across a full range of contexts before moving on to the next rule. Other learners (dubbed "communicators") display control of a rule in only one or two contexts before moving on along the scale." The claims of the Multidimensional model are somewhat weakened by the fact that features seem to be reclassified as variable whenever they prove not to fit the pattern and that trialling in Hawaii showed very wide "profiles with a learner at stage X+4 still producing X+3 features yet managing X+5 ones (Larsen-Freeman and Long 1991:284–8) as well as the fact that the research methodology on which it is based has been severely criticised (Hudson 1993). However, these criticisms reinforce the incomplete nature of current evidence, and the inadvisability of fixing on any "favourite" mistakes to provide a rule of thumb to separate sheep from goats.

To be fair to ACTFL, at least one study has shown that, whatever is said about typical errors in the Guidelines, ACTFL raters do appear in fact to proceed in a more sensible holistic, manner, and that grammatical accuracy is in fact only one aspect taken into account (Magnan 1988).

Even when simplistic rules of thumb are avoided, setting a cut-off to decide the difference between a "2" and a "3" requires a value judgement" (Cronbach 1961:335 cited in Davies 1988:33); "no amount of data collection, data analysis and model building can replace the ultimate judgmental act of deciding which performances are meritorious or acceptable and which are unacceptable or inadequate" (Jaeger 1976:2 cited in Jaeger 1989:492). "Our choice of standard is always a qualitative decision. No measuring system can decide for us at what point "short" becomes "tall". Expert judgement is required." (Wright and Grosse 1993:316). There are lots of ways used to set standards; they tend produce contradictory results and they are difficult to defend because often, as Clark scathingly summarises, they give the impression that they have been "plucked out of the air on the basis of intuition, which is frequently shown on closer examination to be wrongly conceived"(Clark 1987:44).

Both Clark and Stern (1989:214) propose developing norms of performance in real classrooms into definitions of expected

performance, rather than relating them to "some neat and tidy intuitive ideal" (Clark 1987:46). This posits an empirical basis to the development, which can be provided by the Rasch Rating Scale Model. This doesn't alter the fact that the difficulty of a descriptor on the scale will be fixed in relation to a convention in terms of how it is interpreted, but it means that that convention will based upon a relatively wide and consistent consensus,—rather than copied unthinkingly from an existing scale, and that that conventional interpretation will be objectively calibrated. Objectivity is defined as "the requirement that the measures produced by a measurement model be sample free for the agents (test-items) and test-free for the objects (people)" (Wright and Linacre 1987: 2) and the Rasch model offers this characteristic. The scalar Rasch model—for analysing judgements—is called the Rating Scale Model (Wright and Masters 1982). The many-faceted version of the Rating Scale Model (Linacre 1989) has the added attraction, as commented when discussing the requirement for context free—context relevant descriptors, that systematic variation in the subjectivity of different groups in giving scale values to descriptor elements can be investigated.

2.2.2. Number of Levels:

The issue of the number of levels appropriate for a scale is both a pragmatic and an empirical one; the designer must, at the end of the day, decide which principle will dominate. However, it seems common sense to make an informed decision explicitly, rather than being later faced with a choice between (a) revising the scale, thus upsetting a lot of people and reducing its credibility, and (b) living with a scale which has levels which are not used—or which become discredited.

On the one hand, as discussed under Origins of Scales, existing units of time or exit points may dictate a need for (a range of) attainment targets, and in a common scale providing enough steps for all sectors to see progress may lead to a large number of levels (e.g. 20). On the other hand there are clear limits to the number of steps people can distinguish between consistently—which may vary across dimensions—and this can be demonstrated empirically with reliability and separability statistics (either classical statistics like reliability estimates and point-biserials or their Rasch equivalents). "This tension between wanting more (pedagogic) levels to motivate and fewer (natural or critical) levels to establish

equivalencies" (North 1992a:162) needs to be resolved. One solution to this tension would appear to be to develop the scale empirically and then to deduce the number of levels which the statistical properties suggest are a sensible, maximum number of levels. If some people then want to reduce the number of levels for political or operational reasons, there is no reason why they should not do so; reliability should not be effected. If, on the other hand want to increase the number of distinctions further by establishing local "waystages" for continuous, formative assessment between the common levels used for summative assessment (Clark 1984:7), again there is no reason why they should not do so provided they are aware of the distinction.

3. Summary

The issues outlined above are often circumvented rather than addressed in the development of a scale of proficiency. As well as adopting the assumptions of previous scales, some scale writers avoid concrete statements about what learners can do and select categories which they then spread equally across a predetermined range of levels. Distinctions between levels in the statements are then often made by juggling with qualifiers like "some" "a few" "many" "the majority of" etc. The inadequacy of such an approach is acknowledged (Champney 1941, Alderson 1991a). Such descriptors cannot themselves provide criteria for judgements; consistent interpretation becomes impossible unless raters are trained to interpret the descriptors and rate samples of performance in the same way, which they then tend to do without referring to the wording (Jones 1985:77). It is then questionable whether such an approach can be described as criterion-referenced assessment (Skehan 1984:217), and whether raters trained to rate identically produce valid measurement or standardised error (Saal, Downey and Lahey 1980, Wherry 1952, cited in Landy and Farr 1980, 1983). In any case, such an approach is not open for a common framework since, even though one can provide samples of performance, it is difficult if not impossible to avoid people interpreting vague, norm-referenced terminology in relation to the local norms with which they are familiar—rather than in relation to the intended common framework.

A common framework scale should seek to take account of these issues. Put briefly, it should be possible to relate the development of

they are familiar—rather than in relation to the intended common framework.

A common framework scale should seek to take account of these issues. Put briefly, it should be possible to relate the development of the scale to both descriptive theory, and to measurement theory. It should relate to a competence model, yet it should develop a metalanguage and descriptor style which is accessible and relevant to practitioners. It should consciously formulate concrete descriptors of relevant aspects of what people can do in the language, and should empirically establish scale values for these descriptors in relation to the proficiency of relevant groups of learners. In so doing, it should investigate the stability of the scale values for particular descriptors in the different contexts represented by those groups of learners, in order to determine the extent to which the scale can be said to be context-free. This is of course an ambitious undertaking.

4. References

ACTFL (1986) 'ACTFL Proficiency Guidelines'. In (e.g.) Byrnes H. and Canale M. (eds) (1987) *Defining and Developing Proficiency: Guidelines, implementations, and concepts*. Lincolnfield Ill. National Textbook Company

Alderson J.C & North B. (1991) (eds) *Language Testing in the 1990s* Modern English Publications/British Council, Macmillan

Alderson, J.C. (1991a) 'Bands and Scores'. In Alderson J.C & North B:71–86

ALTE (1994) 'ALTE "Can do" statements (draft) drawn on in creating descriptions of levels' in ALTE Document 4 (1994) *The ALTE Framework: A Description of the Framework of the Association of Language Testers in Europe*

Bachman L. (1987) 'Problems in Examining the Validity of the ACTFL Oral Interview' In Valdman A.(ed.) 'Proceedings of the Symposium on the Evaluation of Foreign Language Proficiency'. (Reprinted 1988 in *Studies in Second Language Acquisition* 10/2:149–164)

Bachman, L. (1990) *Fundamental Considerations in Language Testing*, Oxford, OUP

Bachman, L. and Savignon, S.J. (1986) 'The Evaluation of Communicative Language Proficiency: A Critique of the ACTFL Oral Interview' *Modern Language Journal* 70/4:380–90 Win 1986

Barnwell D. (1991) 'Proficiency Testing and the Schools' 1991 *Hispania* 74/1:187–89

Blanche P. (1986) *The Relationships between Self Assessments and Other Measures of Proficiency in the Case of Adult Foreign Language Learners*: Unpublished master's thesis, University of California, Davis California.

Blanche P. (1990) 'Using standardised Achievement and Oral Proficiency tests for Self Assessment Purposes: the DLIFLC Study' *Language Testing* 7/2

Brindley G. (1986) *The Assessment of Second Language Proficiency: Issues and Approaches*. Adelaide. National Curriculum Resource Centre..

Brindley, G. (1989) *Assessing Achievement in the Learner Centred Curriculum* NCELTR Macquarie University Sydney

Brindley G. (1991) 'Defining Language Ability: The Criteria for Criteria'. In Anivan S. (ed) *Current Developments in Language Testing*, Singapore, Regional Language Centre.

Brindley G. (forthcoming) 'Describing Language Development? Rating Scales and Second Language Acquisition', in Bachman L.F. and Cohen A.D. (eds.) *Interfaces between SLA and Language Testing Research*. Cambridge. Cambridge University Press.

Broadfoot P. (1986b)(ed) *Profiles and Records of Achievement: a review of issues and practice*, London Holt, Rinehart and Wilson

Canale M. (1984) 'On Some Theoretical Frameworks for Language Proficiency.' In Rivera, C. (ed) *Language Proficiency and Academic Achievement*, Multilingual Matters

Canale M and Swain M. (1980) 'Theoretical Bases of Çommunicative Approaches to Second Language Teaching and Testing.' *Applied Linguistics* 1/1.1–47

Carroll B.J. (1978/1981) 'Specifications for an English Language Testing Service.' In Alderson and Hughes (1981):66–110

Carroll B.J. (1980) *Testing Communicative Performance* Pergamon, Oxford.

Carroll B.J. and West, R (1989). *ESU (English Speaking Union) Framework. Performance Scales for English Language Examinations.* London: Longman.

Carroll J.B (1967) 'Foreign Language Proficiency Levels Attained by Language Majors near Graduation from College. *'Foreign Language Annals* 1:131–151; cited by e.g. Valette 1991.

Champney, H. (1941) 'The Measurement of Parent Behavior.' *Child Development*, 12, 131–66

Choppin, Bruce (1981) 'Criterion-Referencing of Performance by Latent-Trait Scaling' Paper presented at the *Annual Meeting of the Military Testing Association* October 26–30, 1981

Clahsen, H. (1985) 'Profiling Second Language Development: A procedure for assessing L2 proficiency.' In Hyltenstam and Pienemann

Clark. J.D. (1984) *Language Curriculum in the 1980s.* Babel 19/1.

Clark J.L. (1985) 'Curriculum Renewal in Second Language Learning: an overview.' *Canadian Modern Language Review* 42/2:342–360

Clark J.L. (1987) *Curriculum Renewal in School Foreign Language Learning,* Oxford, Oxford University Press.

Criper C. and Davies A. (1986/1988) *Edinburgh ELTS Validation Project Report.* Reprinted as IELTS Research Report 1 (i), British Counmcil/University of Cambridge Local Examination Syndicate

Cronbach, L. J. (1961) *Essentials of Psychological Testing*, 2nd Edition, London, Harper & Row, cited in Davies 1988

Dade County Board of Public Instruction, (1978), 'Oral Language Proficiency Scale'. *ESOL Placement Interview Guidelines*. ED 179 566

Davies A. (1985) 'Follow My Leader: Is that what language tests do?' In Lee et al 1985: 3–14

Davies, A. (1988) 'Operationalising Uncertainty in Language Testing: An Argument in Favour of Content Validity'. *Language Testing* 5/1:32–48

Davies, A. (1990) *Principles of Language Testing*. Oxford. Blackwell

De Jong H.A.L. (1992) 'Assessment of Language Proficiency in the Perspective of the 21st Century.' In Matter J.F (ed.) Language Teaching in the Twenty-first Century: Problems and Prospects, *AILA Review* 9:39–45

Dickinson L. (1987), *Self Directed Learning*, Oxford, OUP

Douglas D. (1988) 'Testing Listening comprehension in the context of the ACTFL Proficiency Guidelines'. *Studies in Second Language Acquisition* 10/2:245–262

Edwards, V. (1985) 'Assessment of Core French: The New Brunswick Experience'. *Canadian Modern Language Review*; 42/2: 440–51 Nov 1985

Einhorn, H.L. (1974) 'Expert Judgement: Some Necessary Conditions and an Example.' *Journal of Applied Psychology* 59/5:562–571

Ellis R. (1989) 'Are Classroom and Naturalistic Acquisition the Same? A Study of Classroom Acquisition of German Word Order Rules.' *Studies in Second Language Acquisition* 11.

ELTDU (English Language Teaching Development Unit) (1976) *Stages of Attainment Scale* (developed when ELTDU was part of OUP)

Fulcher G. (1987) 'Tests of Oral Performance: the need for data-based criteria'. *ELT Journal* 41/4 287–291

Fulcher G. (1993) *The Construction and Validation of Rating Scales for Oral Tests in English as a Foreign Language*. PhD thesis, University of Lancaster

Galloway, V.B. (1987) 'From Defining to Developing Proficiency: A Look at the Decisions.' In Byrnes H. and Canale M. (eds) (1987 *Defining and Developing Proficiency: Guidelines, implementations, and concepts*. Lincolnfield Ill. National Textbook Company

Glaser, R. (1963). 'Instructional Technology and the Measurement of Learning Outcomes'. *American Psychologist*, 18: 5(19–521

Glass G.V. (1978) 'Standards and Criteria'. *Journal of Educational Measurement* 15, 237–261

Griffin, P.E. (1989) 'Monitoring Proficiency Development in Language'. Paper presented at the *Annual Congress of the Modern Language Teachers Association of Victoria*, Monash University, July 10–11 (1989

Hambleton R.K. (1978) 'Test Score Validity and Standard Setting Methods.' In Berk *Criterion-referenced Measurement*. Johns Hopkins Press

Hargreaves P. (1992): 'Round Table Discussion on the European Language Portfolio'. In *Council of Europe* (1992):150–158.

Hawkey, R. (1980) 'Needs Analysis and Syllabus Design for Specific Purposes'. In Altman and James (eds.) *Foreign Language Teaching, Meeting Individual Needs*, Pergamon

HÉbert Y. (1990) 'The General Language Education Syllabus in Summary.' *The Canadian Modern Languages Review* 47/1.

Henning, Grant (1990) 'Priority Issues in the Assessment of Communicative Language Abilities'. *Foreign Language Annals;* 23/5:379–84 Oct (1990)

Hudson T. (1993) 'Surrogate Indices for Item Information Functions in Criterion-referenced Language Testing.' *Language Testing* 10/2:171–192

Hulstjin, J.H. (1985) 'Testing Second Language Proficiency with Direct Procedures'. A response to Ingram. In Hyltensstam K. and Pienemann M. (eds).

Hyltensstam K. and Pienemann M. (eds) (1985) 'Modelling and Assessing Second Language Development'. Multilingual Matters.

IBM (1974) 'IBM France Performance Charts; Appendix B' in Trim, J.L.M. (1978)

Ingram, D. E. & Clapham, C. (1988) 'ELTS Revision Project: A New International Test of English Proficiency for Overseas Students'; Paper presented at the *Combined Annual World Congress on Language Learning of the Federation Internationale des Professeurs de Langues Vivantes* (16th) and *Biennial National Languages Conference of the Australian Federation of Modern Language Teachers Associations* (7th) Australian National University, Canberra, Australia, January 4–8, (1988)

Ingram D. E. and Wylie E. (1989) 'Developing Proficiency Scales for Communicative Assessment'. Paper given at the National Assessment Consultation for the National Assessment Framework for Languages at Senior Secondary Level, (NAFLaSSL) Sydney, 5 December (1989).

Jaeger R.M. (1976) 'Measurement Consequences of Selected Standard Setting Models'. Florida Journal of Educational Research, cited in Jaeger (1989):492.

Jaeger R.M. (1989) 'Certification of Student Competence' in Linn R.L. (ed) *Educational Measurement* 3rd edition American Council on Education/Macmillan, New York.

Jones N. (1993) 'An Item Bank for Testing English Language Proficiency: using the Rasch model to construct an objective measure.' PhD thesis, University of Edinburgh.

Jones R.L. (1985) 'Some Basic Considerations in Testing Oral Proficiency'. In Lee et al (1985) 77–84

Kohonen, V. (1989). 'Evaluation in Relation to Learning and Teaching of Languages for Communication'. Paper read at the *Council of Europe Symposium* in Sintra, Portugal.

Kramsch C. (1986) 'From Language Proficiency to Interactional Competence'. *Modern Language Journal* 70/4: 366–72

Ladousse G.P. (1982) 'From Needs to Wants: Motivation and the Language Learner' *System* 10/1:29–37

Landy, F.J. & Farr, J. (1983) *The Measurement of Work Performance* San Diego, CA. Academic Press.

Lantolf, J. and Frawley W. (1985) 'Oral Proficiency Testing: A Critical Analysis'. *Modern Language Journal* 70: 337–345

Lantolf, J. and Frawley W. (1988) 'Proficiency, Understanding the Construct'. *Studies in Second Language Acquisition* 10/2: 181–(196

Lantolf, J. and Frawley W. (1992) 'Rejecting the OPI—Again. A Response to Hagen'. *ADFL Bulletin* 23/2 34–37.

Larsen-Freeman D, and Long. M. (1991) *An Introduction to Second Language Acquisition Research.*, Harlow, Longman

Lee Y.P., Fok C.Y.Y, Lord R. and Low G. (eds) (1985) *New Directions in Language Testing*, Oxford, Pergamon

Linacre J.M. (1989) *Multi-faceted Measurement* Chicago, MESA Press.

Linacre J.M (1991) 'Beyond Partial Credit, Rasch Measurement', *Transactions of the Rasch Measurement Special Interest Group of the American Educational Research Association*, 5/2 Summer (1991: 155

Linacre J.M., Engelhard G., Tatum D.S. and Myford C. (1992) 'Measurement with Judges: Many-faceted Rasch Analysis'. Unpublished paper

Liskin-Gasparro, J. E. (1984) 'The ACTFL Proficiency Guidelines: Gateway to Testing and Curriculum'. *Foreign Language Annals;* 17/5: 475–89) Oct (1984)

Lowe, P. (1983) 'The ILR Oral Interview: Origins, Applications, Pitfalls, and Implications'. *Unterrichtspraxis;* 16/2: 230–44 Fall (1983)

Luoma S. (1993) 'Validating the (Finnish) Certificates of Foreign Language Proficiency'. Paper presented at the 15th *Language Testing Research Colloquium,* Cambridge and Arnhem, 2–4 August (1993)

Mager R.F. (1962) *Preparing Instructional Objectives,* Palo Alto, CA: Feardon Publishers, cited in Glass (1978).

Magnan, S. S. (1988) 'Grammar and the ACTFL Oral Proficiency Interview: Discussion and Data'. *Modern Language Journal;* 72/3: 266–76) Fall (1988)

Maley, A. (1980) 'Realism and Surrelaism in foreign Language Teaching': *Récherches et Echanges* 5/2:1–8

Matthews M. (1990a) 'The Measurement of Productive Skills. Doubts Concerning the Assessment Criteria of Certain Public Examinations'. *ELT Journal* 44/2: 117-120

McNamara T. (1996) *Language Performance Assessment.* London Longman

Milanovic M., Saville, N, Pollitt, A. & Cook A. (1992) 'Developing and Validating Rating Scales for CASE: Theoretical Concerns and analyses.' Paper presented at the 14th Annual *Language Testing Research Colloquium*

Morrow K (1977). *Techniques of Evaluation for a Notional Syllabus.* Royal Society of Arts, London

Morrow, K. (1986) 'The Evaluation of tests of Communicative Performance'. In Portal M (1986) *Innovations in Language Testing* NFER-Nelson Windsor

Munby J. (1978) *Communicative Syllabus Design*, Cambridge CUP

Murphy K.R. and Cleveland J.N. (1991) *Performance Appraisal: An Organisational Perspective*. Boston: Allyn and Bacon.

Murphy, K.R. and Constans J.I. (1987) 'Behavioural Anchors as a Source of Bias in Rating'. *Journal of Applied Psychology* 72/4 573–577

Murphy, K.R., Martin, C. & Garcia, M. (1982). 'Do Behavioral Observation Scales Measure Observation?' *Journal*

Murphy, K.R. and Pardaffy V.A. (1989) 'Bias in Behaviourally Anchored Rating Scales: Global or Scale Specific.' *Journal of Applied Psychology* 74/2, 343–346

North B. (1991) 'Standardisation of Continuous Assessment Grades'. In Alderson & North: 167–177

North B. (1992) 'A European Language Portfolio: Some Options for a working Approach to Design Scales for Proficiency'. In Council of Europe (1992): 158–174; reprinted in Schärer R. & North B. 'Towards a Common European Framework for Reporting Language Competency', *NFLC Occasional Paper*, National Foreign Language Center, Washington D.C., April (1992)

North B (1993) L'Evaluation Collective dans les Eurocentres, in Evaluations et Certifications en Language Etrangére, numÉro spÉcial, Le Français dans le Monde—RÉcherches et Applications, (1993): 69–81

North B. (1993) *Scales of Language Proficiency, A Survey of Some Existing Systems*, Strasbourg, Council of Europe

North B. (1993) 'The Development of Descriptors on Scales of Proficiency: perspectives, problems, and a possible methodology' *NFLC Occasional Paper*, National Foreign Language Center, Washington D.C., April (1993)

North (1994) 'Itembanker: A Testing Tool for Language Teachers'. *Language Testing Update* 16, Autumn (1994) : 85–97

Nuttall D. and Goldstein G. (1986) 'Profiles and Graded Achievements: the technical issues.' In Broadfoot P. (ed) :183–202 (ed) *Profiles and Records of Achievement: a review of issues and practice,* London Holt, Rinehart and Wilson

Oscarson M. (1978/9) *Approaches to Self Assessment in foreign Language Learning:* Strasbourg, Council of Europe (1978; Oxford, Pergamon (1979.

Oscarson, M. (1984) *Self Assessment of Foreign Language Skills: a survey of research and development work.* Strasbourg, Council of Europe

Oscarson, M (1988) 'Self Assessment of Communicative Proficiency'. In Trim J.L.M. et al. *Evaluation and Testing in the Learning and Teaching of Languages for Communication* Council of Europe.

Page B. (1992) 'Some Fundamental Issues in Designing a Framework'. In *Council of Europe* (1992): 135–139

Parks M.R. (1985) 'Interpersonal Competence and the Quest for Personal Competence.' In M.L. Knapp and G.L. Miller *Handbook of Interpersonal Communication,* Beverley hills, Calif. Sage.

Pienemann M. and Johnson M. (1987). 'Factors influencing the Development of Language Proficiency.' In Nunan D. (ed) *Applying Second Language Acquisition Research.* Adelaide, National / Curriculum Resource Centre.

Pollitt A. (1991) 'Response to Alderson, Bands and Scores'. In Alderson & North: 87–94

Pollitt A. and Hutchinson C.(1987) 'Calibrating Graded Assessments; Rasch Partial Credit Analysis of Performance in Writing' *Language Testing* 4: 72–92

Pollitt A. and Murray N.L. (1993) 'What Raters Really Pay Attention to'. Paper presented at the 15th *Language Testing Research Colloquium,* Cambridge and Arnhem, 2–4 August (1993

Popham W.J. (1978) *Criterion-Referenced Measurement,* Prentice Hall, Englewood Cliffs, N. J., cited in Hambleton (1978

Quellmalz, E. (1982a) 'Scale for Evaluating Narrative Writing' ED 236 653

Quellmalz, E. (1982b) 'Scale for Evaluating Expository Writing' ED 236 670

Raffaldini T (1988) 'The Use of Situation Tests as Measures of Communicative Ability'. *Studies in Second Language Acquisition* 10/2 (197–216

Richterich, R. (1983) 'Introduction' in Richterich (ed.) *Case Studies in Identifying Language Needs*, Pergamon Press

Robinson, P. (1983) 'ESP, Communicative Language Teaching and the Future.' In Johnson and Porter (ed) *Perspectives in Communicative Language Teaching*, Academic Press

Romiszowski, A.J. (1981) *Designing Instructional Systems*, Kogan Page

Saal F.E. Downey R.G. and Lahey M.A (1980) 'Rating the Ratings: Assessing the Psychometric Quality of Rating Data'. *Psychological Bulletin* 88/2 413–428

Savignon, S.J. (1985) 'Evaluation of Communicative Competence: The ACTFL Provisional Proficiency Guidelines'. *Canadian Modern Language Review*, 41/6 May (1985: 1000–1007 (reprinted also in Modern Language Journal 69. 129–34)

Schneider G. and Richterich R. (1992) 'Transparency and Coherence: Why and for Whom?' In *Council of Europe* (1992): 43–49

Schärer R. (1992) 'A European Language Portfolio—a Possible Format'. In *Council of Europe* (1992): 140–146

Shohamy, E. (1981) 'Inter-rater and Intra-rater Reliability of the Oral Interview and Concurrent Validity with Cloze Procedure'. In Palmer , A.S. Groot, J.M., Trosper, G.A. (1981) *The Construct Validation of Tests of Communicative Competence.* TESOL (1981

Skehan P. (1984) 'Issues in the Testing of English for Specific Purposes'. *Language Testing* 1,2, 202–220

Skehan P. (1989) *Individual Differences in Second and Foreign Language Learning*. Edward Arnold, London,

Skehan P. (1995) 'A Framework for the Implementation of Task-based Instruction', *Applied Linguistics* 16/4: 542–566.

Skilbeck M. (1982) 'Three Educational Ideologies'. In Horton T. and Raggat P. *Challenge and Change in the Curriculum*, Sevenoaks, Hodder and Stoughton, cited in Clark J.L. (1985); (1987)

Smith, K. (1992) 'Self Evaluation in Foreign Language Learning', *IATEFL Testing Newsletter* April (1992):4–5

Smith, P. C. & Kendall J.M. (1963) 'Retranslation of Expectations: An Approach to the Construction of Unambiguous Anchors for Rating Scales'. *Journal of Applied Psychology*, Vol 47/2.

Spolsky B. (1986) 'A Multiple Choice for Language Testers'. *Language Testing* 3/2: 147–158

Spolsky B. (1989) *Conditions for Second Language Learning*, Cambridge, Cambridge University Press

Spolsky B. (1993) 'Testing and Examinations in a National Foreign Language Policy'. In Sajavaara, K, Takala S., Lambert D. and Morfit C. (eds) *National Foreign Language Policies: Practices and Prospects*. Institute for Education Research, University of Jyväskylä: 194–214

Stern H.H. (Allen P. and Harley B. eds) (1992) *Issues and Options in Language Teaching*, Oxford, Oxford University Press

Stratton N. (1986) 'Recording Achievement: the City and Guilds Experience', in Broadfoot P. (ed): 109: 126

Swain M. (1992) 'Using Assessment Information in French Immersion Programs', in Shohamy and Walton (1992:73–85

Thorogood J. (1992) *Continuous Assessment and Recording*. Pathfinder 13, CILT

Trim, J.L.M. (1978) *Some Possible Lines of Development of an Overall Structure for a European Unit/Credit Scheme for Foreign Language Learning by Adults,* Council of Europe

Trim J.L.M. (1984) Extract from 'Developing a Unit/Credit Scheme of Adult Language Learning'. Reprinted in Van Ek J. and Trim J.L.M. (eds) *Across The Threshold,* Oxford, Pergamon: 9–26; cited in Douglas (1988)

Valette, R. M. (1991) 'Proficiency and the Prevention of Fossilization—An Editorial'. *Modern Language Journal;* 75/3: 325–28 Fall (1991)

Valette R.M. and Disick R.S. (1972) *Modern Language Performance Objectives and Individualisation: A Handbook* New York, Harcourt, Brace Jovanovich, cited in Stern (1992)

Van Ek, J.A. (1986) *Objectives for Foreign Language Teaching, Volume I: Scope* Council of Europe

Van Els, T. (1992) 'Revising the Foreign Language Examinations of Upper Secondary Education in the Netherlands'. *In Council of Europe* (1992: 109–115)

Van Lier. L. (1989) 'Reeling, Writhing, Fainting and Stretching in Coils: Oral proficiency interviews as conversation.' *TESOL Quarterly* (1989) 489–508

Wesche M. Sima Paribakht and Ready D. (1993) 'A Comparative study of four Placement Instruments'. Paper presented at the 15th *Language Testing Research Colloquium,* Camridge and Arnhem, 2–4 August (1993)

Wherry R.J (1952) The Control of Bias in Rating: A Theory of Rating. Unpublished paper: Personnel Board Report 922, Department of the Army, Personnel Research Section, Washington DC, cited in Landy and Farr, (1980), (1983).

Widdowson H.G. (1990. *Aspects of Language Teaching.* Oxford. OUP

Wilds C.P. (1975) 'The Oral Interview Test'. In B. Spolsky and R. Jones *Testing Language Proficiency* Washington D.C., Center for Applied Linguistics: 29–44

Wright B.D and Grosse M. (1993) 'How to Set Standards, Rasch Measurement,' *Transactions of the Rasch Measurement Special Interest Group of the American Educational Research Association,* 7/3 Autumn (1993: 315–6

Wright B.D and Linacre J.M (1987) 'Research Notes, Rasch Measurement', *Transactions of the Rasch Measurement Special Interest Group of the American Educational Research Association,* 1/1: 2–3

Wright B.D. & Masters G. (1982) *Rating Scale Analysis: Rasch Measurement* Chicago, Mesa Press