Acknowledgments

I would like to express my thanks first of all to Neil Jones for introducing me to the delights and frustrations of Rasch Modelling. At that time Neil, now working for the University of Cambridge Local Examinations Syndicate (UCLES), was developing *Itembanker* for Eurocentres whilst doing a PhD at Edinburgh with Alistair Pollitt. Thanks also to Dorry Kenyon at the Center for Applied Linguistics in Washington for pointing me in the direction of Mike Linacre and FACETS analysis when I came knocking on his door in 1992 asking: "Isn't there some connection between scales of proficiency and Rasch scaling?" Very many thanks to Mike Linacre (Chicago: MESA Laboratory) and to Alistair Pollitt (now also UCLES) for all the help and advice they themselves provided with what turned out to be a rather complicated analysis.

Thanks to Peter Skehan, then running Thames Valley University ELT Department for his support, guidance and consideration as my PhD supervisor, and to Günther Schneider of the University of Fribourg, who "bought" the idea of this methodology and was an ideal friend and colleague. I would also like to thank the 100 English teachers who let themselves be cajoled into taking part in the 1994 survey which is reported—most of all my wife Leslie for not only taking part but for continuing to talk to me during the following few years when I was, shall we say, distracted.

Finally, I would also like to take this opportunity to thank Dick Lambert (formerly National Foreign Language Center) and the NFLC for offering me the Mellon visiting fellowship in 1992 which enabled me to broaden my horizons, find out what had been happening in terms of scales and scaling outside the language teaching world, and made it feasible for me to give an academic basis to my involvement in the Council of Europe Framework development. Thanks also to the Eurocentres Foundation for their sponsorship of my involvement in the development of the Framework, and for allowing me to go part-time in order to undertake the study reported in this book.

Contents

Figures

Tables

Preface

This study is related to steps in Europe towards the development of a Common Framework for language teaching and learning. The project here described was undertaken primarily to serve two functions:

Firstly to provide an empirical basis in a multi-lingual, multi-sector setting to the development of an illustrative bank of descriptors. These descriptors define at an ascending series of levels various aspects of proficiency related to the parameters specified in the Council of Europe Framework subsequently published in 1996.

Secondly to test in a pilot project for English a methodology to establish the difficulty level of descriptors and the ability level of learners on the same mathematical scale. This was in order to provide transparent descriptors of the foreign language proficiency of learners in the different sectors of the Swiss educational system.

Put another way, the aim of the study was to refine a set of criterion descriptors of English language proficiency through workshops with teachers and then to calibrate those descriptors in relation to assessments by teachers of their learners, in order to provide a basis for the development of instruments for criterion-referenced teacher- and self-assessment in relation to a common reference scale.

After the Introduction, Chapter 1 discusses existing scales of language proficiency and some of the particular problems involved in the development of a common framework scale. The vast majority of existing scales of language proficiency have been developed intuitively by a single author or small committee. Whilst one should not necessarily underestimate the effectiveness of such methods in relation to the development of a scale of definitions for a particular function in relation to particular assessors and particular learners with whom the authors are intimately familiar, the limited validity of such an approach is more obvious in relation to the development of a common framework scale.

The validity of a scale of proficiency is increased if relevant theory is taken into account in its development. Chapter 2 therefore considers linguistic issues, discussing the relationship of competence to proficiency and the relevance of models of language competence, and puts forward an approach to the description of language activity. Chapter 3 then considers measurement issues and the requirements of a valid scale, discusses relevant experience outside the language field, and puts forward an adaptation of itembanking methodology through the application of a scalar version of the Rasch probabilistic measurement model to the calibration of descriptors as items. It also outlines how the use of the many-faceted version of the Rasch scalar model, which can measure and adjust for rater severity, can be used to scale learner achievement in relation to those descriptor items.

Chapter 4 describes the procedures adopted in the project and outlines the way in which the contents of existing scales were analysed into a classified descriptor pool of approximately 2,000 items relevant to spoken interaction. These selected, edited descriptors were then pre-tested in a series of workshops with teachers to refine and reduce the pool to approximately 400 transparent, useful and relevant items. Chapter 5 then describes the organisation of the main data collection for the calibration of the descriptors on the basis of the assessments by 100 teachers of 945 learners of English. The adaptation of an itembanking methodology to data produced by teacher judgements covering the full range of foreign language proficiency is not without its problems and the most important of the complications which arose in the analysis are also discussed in Chapter 5.

Chapter 6 introduces the program used in the analysis—FACETS, by Mike Linacre—and outlines the various steps taken in the analysis and scale construction. The results are discussed in Chapters 7 and 8. Chapter 7 outlines the way in which the arithmetical scale of calibrated items was turned into a proficiency scale of 10 levels through the identification of cut-off points between levels. The coherence of the content thus calibrated at each level is discussed, together with what the calibrated scale appears to say about the development of the proficiency of the learners involved. Chapter 8 discusses the process by which the learners themselves were calibrated onto the same scale and presents provisional conclusions for the ranges of proficiency achieved by learners in the different educational sectors in relation to learning time.

Finally, Chapter 9 draws conclusions and makes recommendations for future research.

Abbreviations

ACTFL	American Council on the Teaching of Foreign Languages
ALTE	Association of Language Testers in Europe
ASLPR	Australian Second Language Proficiency Ratings
c&h	Carroll B.J. and Hall P.J Interview Scale 1985
carr	Carroll B.J. Oral Interaction Assessment Scale 1980
CCSE	University of Cambridge/RSA Certificates in Communicative Skills in English 1990
DIF	Differential Item Functioning
Dutch	Netherlands New Examinations Programme 1992 (Van Els 1992)
EC	Eurocentres Scale of Language Proficiency 1993
EFL	English as a Foreign Language
ELTS	English Language Testing Service
elviri	Elviri et al: Oral Expression 1986 (in Van Ek 1986)
ESL	English as a Second Language (used interchangeably with EFL in a North American context)
ESU	English-speaking Union Framework Project
EurLon	European Certificate of Attainment in Modern Languages 1993
finn	Finnish Nine Level Scale of Language Proficiency 1993
FSI	Foreign Service Institute Absolute Proficiency Ratings
fulch	Fulcher: The Fluency Rating Scale 1993
Got	Goteborgs Univeritet: Oral Assessment Criteria
Hof	Hofmann: Levels of Competence in Oral Communication 1974
IELTS	International English Language Testing Service
ILR	Interagency Language Roundtable Language Skill Level Descriptors
llb	British Languages Lead Body: National Language Standards 1993

Lon	University of London School Examination Board: Certificate of Attainment 1987
NatC	English National Curriculum: Modern Languages 1991
North	Eurocentres: European Language Portfolio Mock-up: Interaction Scales 1991
OTESL	Ontario ESL Oral Interaction Assessment Bands
RSA	Royal Society of Arts Modern Languages Examinations: French 1989
sho	Hebrew Oral Proficiency Rating Grid 1981
Trim	Trim: Possible Scale for a Unit/Credit Scheme: Social Skills 1978
UCLES	University of Cambridge Local Examinations Syndicate
wilk	Wilkins: Proposals for Level Definitions for a Unit/Credit Scheme: Speaking 1978

Introduction

The study described in this book was undertaken as a PhD as part of a research project in Switzerland linked to the development of the Council of Europe Common European Framework for language learning and teaching (Trim 1978; Council of Europe 1992; North et al 1992; Council of Europe 1996). The Swiss research project aimed to identify the range of language competence achieved at points at which learners switch educational sectors (Schneider and North 1999) and to develop a prototype "Language Passport" or "Language Portfolio" to record that achievement (Schärer and North 1992; Schärer ed. 1999). These initiatives were a follow up to the Council of Europe Intergovernmental Symposium hosted by Switzerland at Rüschlikon near Zürich in November 1991 entitled: "*Transparency and Coherence in Language Learning in Europe: objectives, evaluation and certification.*" The Swiss research project sought to make transparent teachers' "fuzzy" internalised norms and standards (Murphy and Cleveland 1991) and to develop them into definitions of expected performance levels for the learners concerned (Stern 1989: 214) in criterion statements presented both in scales of proficiency and in checklist form (Brindley 1989: 62–84), so that they could be used as an information network for criterion-referenced assessment by the partners in the language teaching and learning process (Richterich and Schneider 1992). This was an ambitious aim, because it entailed creating a common framework of reference which will be relevant to different language regions (German-speaking, French-speaking, Italian-speaking) and different educational sectors (lower secondary, upper secondary, vocational, adult) as well as, ultimately, different users (teachers, learners, employers).

The Common European Framework

The aim of the Council of Europe Common European Framework project could be paraphrased as being to develop a framework of reference which will (a) help the different partners in the language teaching and learning

process focus on and plan for the development of the capacities necessary for effective language use at different levels of achievement, (b) enable different providers of language teaching services to describe and compare their systems in terms of a common metalanguage, and (c) provide descriptors of communicative language proficiency in as many of the categories as prove to be scaleable at an ascending series of levels (Council of Europe 1996). The members of the Council of Europe project team hope that the development of the Common European Framework will help learners to plot their progress in terms of recognised benchmarks, help teachers, course writers and educational administrators articulate their concept of communicative language learning, and contribute to improved effectiveness at educational cross-over points—when learners change educational sectors or move house.

Both the Council of Europe and the Swiss research projects thus involved attempts to relate statements about learner achievement in terms of communicative language *proficiency* to theoretical models of communicative language *competence*. The relationship between competence and proficiency is complex, and is related to the distinction between theoretical models and operational models. Scales of language proficiency can be seen as a branch of Behavioural Scaling, the classic methodology for which was developed by Smith and Kendall (1963). Smith and Kendall were reacting to a situation in which scales used in hospitals designed by applied psychologists for rating theoretically-based dimensions were often unworkable and at best unreliable since the "front line" practitioners observed behaviour (rather than reflecting about underlying dimensions) and thought in quite different categories to psychologists. McNamara (1995) has drawn attention to a similar gulf in language testing between operational proficiency approaches, many of which started as pragmatic innovations at a time when the theory of language learning was still structuralist, and theories of communicative competence since developed by, for example, Canale and Swain (1980). Because such models tend to maintain Chomsky's (1965; 1975) distinction between competence and performance, and because like Chomsky they tend to associate competence with knowledge, they continue to have limited applicability for practitioners interested in language use. Skehan (1995a: 16) considers that it is in fact "misconceived to see competence as underlying performance in any straightforward manner" and proposes *ability for use* as something separate from both competence and performance which is itself related to

Bachman's (1990) interpretation of strategic competence. This rather complex discussion is taken further in Chapter 2. The relevant point is that a Common Framework needs to strike a balance between categories which can be justified from the incomplete state of theory, and categories which can be understood by and felt to be relevant by the users.

Scaling Proficiency Descriptors

Both the Council of Europe and the Swiss research projects also involved attempts to assign descriptions of language ability to one level or another—that is to *scale* descriptors. Considering the extent of the literature on scaling, (e.g. Thorndike 1904/1916; Freyd 1923; Thurstone 1928a, 1928b; Champney 1941; Rasch 1960/80; Wright and Stone 1979; Wright and Masters 1982) and on behavioural scaling in particular (e.g. Smith and Kendall 1963; Landy & Farr 1983; Borman 1986), it is in fact surprising how little use has been made of scaling theory or of empirical development in the production of language proficiency *scales*. Virtually all language scales seem to have been produced on the basis of intuition and /or subjective matching to samples of performance by a small authoring team (c.f. North 1993a for reviews). Whilst this approach appears to be successful in systems for particular examinations or institutions dealing with a specific kinds of learners and specific kinds of raters which can, to some extent, impose a common interpretation of the wording, it stands to reason that such an approach is likely to be less successful in a common framework since no one has a sufficient overview of the way different types of teachers may interpret the same wording in relation to different types of learners. It is difficult, in other words, to move from the specific to the general.

Aim of the Study

The aim of this study was to develop an example set of descriptors of communicative language proficiency which (a) bear some relation to the theory-based categories used for the description of communicative language competence in the Council of Europe Common European Framework, which (b) built explicitly on collective experience in the field of scales of language proficiency, which (c) were felt to be clear, comprehensible and relevant by practising teachers, and which (d) were calibrated with a measurement model in relation to the achievement of learners in different educational sectors in a multi-lingual context. The measurement scale and

the set of calibrated descriptors produced were used as the basis for an illustrative set of scales of communicative language proficiency for the Common Reference Levels in the Council of Europe Common European Framework (Council of Europe 1996). The study also served to develop and pilot the qualitative and quantitative methodology for the development and calibration of descriptors and learner achievement in the Swiss research project. The methodology was repeated in a larger follow-up Swiss survey in 1995 which includes French and German as foreign languages as well as replicating the study for English (Schneider and North 1999).

Methodology

The methodology for this study followed four steps: (a) analyse existing scales of language proficiency in terms of categories which can be related to theories of language and to the emerging Council of Europe Common European Framework model, and write descriptors for those aspects of proficiency which appeared to be under-represented; (b) reduce and refine the descriptor set using groups of teachers as informants; (c) calibrate the descriptors felt to be the clearest and most relevant through an analysis of the judgements of Swiss teachers using a scalar version of the Rasch model from the Item Response Theory family of measurement models (Wright and Masters 1982, Linacre 1989), and finally, (d) create a common scale of language proficiency from the resultant hierarchy of descriptors (Griffin 1989, 1990a, 1990b).

The Context of the Study

The study took place within the context of moves towards a common European metric for recording language competence to calibrate the current "proliferation of rating scales" (De Jong 1992: 43). This search for a common metric is, as Bachman comments, a common thread running through much literature on language testing (Bachman 1990a: 5). In the US, the need to maintain the momentum towards such a common metric "with all of its difficulties" has been described as "essential for the development of a meaningful national language policy in foreign language learning and use" (Lambert 1993a: 155) and such a metric has been also proposed as a means of comparing research outcomes (Clark and O'Mara 1991: 83). In Europe the motivation for a common framework (or metric) is more pragmatic, more socio-political than academic. In a world of greatly

increased international communications and trade, personal mobility and the interrupted schooling and lifelong learning which that entails, such a framework (Trim 1978, 1992) would facilitate comparison between systems (Carroll & West 1989; Carroll 1992), describe achievement in terms meaningful to teachers, learners, parents and employers (Richterich and Schneider 1992), and allow learners to plot a personal record of achievement during their language learning career across different institutions (Schärer 1992).

Switzerland as a Test-bed

Switzerland is in many respects an ideal test-bed for experimentation with such developments. First of all the country is governed in a very decentralised and even relatively informal manner, is small enough for people involved in applied linguistics to all know one another, and does not have rigid demarcation lines between the state and private sectors. Secondly the educational system is equally decentralised. Each of the 26 cantons is sovereign in this respect. There is no department of education and there are no powerful examination boards. Finally and most fundamentally Switzerland is a multi-lingual country. There are four language regions (German-speaking, French-speaking, Italian-speaking, Romansch-speaking). The French- and German-speaking cantons have pedagogic cultures which are quite similar to those of neighbouring states speaking the same languages, which gives the Swiss educational system a distinct pluralism. The population also includes over 25% foreign nationals, the vast majority of whom speak a mother tongue other than the four national languages, and in addition the majority of educated people in the country understand English (as well as speaking a second national language) even if they do not necessarily admit to being able to remember how to speak it.

Switzerland, though small and somewhat unique, was therefore sufficiently complex and sufficiently international to offer a sort of model of the situation at a European level. Switzerland offers the opportunity to develop a common framework scale taking account of different educational sectors, different language regions and pedagogic cultures, and different mother-tongues. Establishing validity in contexts which differ by these demographic variables is of central concern to the development of any common framework. A second advantage of the Swiss context for such research is that the concepts of scales of language proficiency and of descriptors of proficiency were almost totally novel to some 90% of the

teachers who took part in the survey. Therefore it was possible to investigate the interpretations of teachers from different contexts to the same formulations without those reactions being significantly coloured by familiarity with either those particular descriptors or the concept of descriptors.

Perhaps because of this novelty, Swiss teachers of English in fact proved to be receptive to the ideas represented by this study. One of the perceived disadvantages of the decentralised nature of the current system is the difficulty of relating educational achievement in one sector or canton to achievement in another. With the exception of the pre-university Matura/Maturité qualification, assessment is based almost entirely on teacher judgements expressed in grades like "4.5" (out of a possible 6.0). These grades are norm-referenced: given in relation to what the teacher considers to be a reasonable achievement for that particular class. If a student changes class, the grades given change dramatically. Future prospects for young people not going on to university are determined by a balance between these grades and the impression of a potential employer on a week's trial apprenticeship. The system works well in terms of what appears to be the primary aim of the Swiss educational system—ensuring that as many children as possible are accepted for apprenticeships—but it lacks transparency. In terms of foreign language teaching the lack of transparent objectives and a common metalanguage for expressing achievement is perceived by the Standing Conference of Cantonal Directors of Education as reinforcing an apparent tendency for teachers in each successive educational sector to undervalue the foreign language achievement of the individual children who come to them from the lower educational sector. This is thought to have serious negative effects on the motivation of the children concerned, and on that of the teachers of the feeder educational sector.

A European Language Portfolio

One response in Switzerland to this problem is the proposed European Language Portfolio. This is a personal document, the property of the learner. One section, called the Language Passport, profiles proficiency in different foreign languages in terms of a *common European scale*. Since Switzerland is a multi-lingual society with three large neighbours, however, there does not seem to be much point to such a passport if recognition of it is confined to Switzerland. This is one reason why Switzerland hosted the

Council of Europe Symposium "*Transparency and Coherence in Language Learning in Europe: objectives, evaluation and certification*" in 1991, which was primarily concerned with discussion of the development of a Common European Framework of reference, but at which a European Language Portfolio was put forward as a Swiss proposal (Schärer 1992; North 1992a). This proposal was unanimously accepted by the Symposium, the formal text reading as follows (my italics):

> "The Portfolio should contain a section in which formal qualifications are related to a *common European scale*, another in which the learner him or herself keeps a personal record of language learning experiences and possibly a third which contains examples of work done. Where appropriate, entries should be situated within the Common Framework." (Council of Europe 1992: 40)

A recent report on the development and trialling of versions of the Portfolio for different educational sectors in an international Council of Europe network is given in a special edition of *Babylonia*, (Schärer, ed. 1999). The first concrete steps towards the development of a Portfolio were taken in November 1993 when the Swiss National Science Research Council funded a project to identify and describe the range of language proficiency achieved at points at which learners switch educational sectors. This research project (Schneider and North 1999), referred to at the beginning of this introduction, was the context for the study described in this book. The pilot investigation which is the subject of this book was conducted for English because the aim was to create a common scale, and English is the only language taught widely in all cantons.

The identification of learner achievement at the end of educational sectors was, however, only one aim of the Swiss research project. The descriptors calibrated in this study have been exploited to develop continuous assessment checklists for teacher and learner use. The aim of producing such instruments related to a common framework scale is not just to promote continuous teacher assessment with transparent criteria, but also to promote learning by developing metacognitive skills, boosting self confidence and thus increasing motivation since "studies looking into the motivational patterns of school learners have established that in school foreign language learning motivation is best generated by a feeling of successful achievement" and "in a language learning context nothing succeeds like success" (Clark 1987: 75–7). Whilst, as Skehan points out

(1991a: 281), there is some disagreement over whether success causes motivation (Burstall 1975: 13) or motivation causes success (Gardner 1985), the former appears at least as likely as the latter (McLaughlin 1987: 126) and Gardner himself proposes: "in this model…motivation is shown not as a direct cause of achievement but rather as an indirect cause through self confidence" (Gardner 1985: 164). A further objective of developing such a framework of a common scale and related instruments was the opportunities for teacher development which involvement in such a project entails. In his attempt to evaluate the English schemes for graded objectives in modern languages (GOML) Harrison (1982a, cited in Nuttall and Goldstein 1986: 194) concluded that while lack of formal evaluation data made conclusions difficult to draw, the key ingredients in the success of the schemes (seen in vastly increased take up of advanced courses) was probably due to (a) the positive reinforcement at regular intervals, but also (b) the enthusiasm of the teachers.

A first step towards such a system, however, is the development of a common scale of language proficiency and it is with this aspect of the project that this book is concerned.

1 Language Proficiency Scales

Scales of language proficiency have become relatively widespread over the past decade as part of a general movement towards more transparency in educational systems, which 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 1990s saw quite a proliferation with, for example, the British National Language Standards (Languages Lead Body 1992), the Eurocentres Scale of Language Proficiency (North 1993c), the Finnish Scale of Language Proficiency (Luoma 1993) and the ALTE Framework (Association of Language Testers in Europe 1994). Many of these scales represent what Bachman (1990: 325–330) has described as the "real-life" 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 1992/6; Fulcher 1993; Upshur and Turner 1995; Brindley 1998).

The following extract from the mid range of the 10 band Eurocentres global scale is a fairly typical 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 different purposes. 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.

> 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.*

That this particular scale *does* fulfil its purpose is suggested by a 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 small-group interaction in the classroom: 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 (e.g. Oscarson 1984; Blanche 1986, 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.

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 does not 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 1998: 22).

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; Richterich and Schneider 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 scored 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 above, scales of language proficiency have the potential to exert a positive influence on the orientation, organisation and reporting of language learning.

Origins

The many names given to scales of language proficiency 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, or as stages of attainment.

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 Carroll 1980; Shohamy 1981; Morrow 1977, 1986; Alderson 1991; Milanovic et al 1992/6).

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:

Firstly, 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).

Secondly, 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," 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 heavily relative wording. For example in the oral interviews which form Paper 5 of the First Certificate in English (FCE) and the Certificate of Proficiency in English (CPE), a rating scale is used to grade the candidate in relation to the pass norm for the examination in question. For 10 years (1984–94) both tests used very similar rating scales, and it seems to be a feature of such rating scales that you need to be virtually a native speaker to get the top grade, or a beginner to get the bottom grade, whatever level the examination is. Rating scales revised in the late 1990s for the full Cambridge examination suite continue this norm-referencing. Similar wording of the top and bottom ranges on the scales for different examinations are intended to mean something different in the context of each exam.

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). 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 Considerations (in effect interlocutor support). Once again, although the philosophy is pass/fail mastery and although the grid is 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. An 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). Work on the development of a series of "Can do" descriptors for the levels is currently being undertaken.

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 Eurocentres aim); school years (original English National Curriculum aim).
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.

It is a feature of scales of this type 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 Eurocentres 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 synthesizing 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).

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). Bachman and Savignon (1986) and Fulcher (1993/6) suggest that doing so limits the generalisability of the test result to similar tasks and situations.

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 qualita-

tive 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:

What the learner can do:	- user-oriented	(simpler)
	- constructor-oriented	(more developed)
How well he/she performs:	- assessor-oriented	(simpler)
	- diagnosis-oriented	(more developed)

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 (*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 to 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 (*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 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. (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 analysis 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. For these reasons, the needs analysis approach has not been applied in this study, though successive groups of teachers were asked in a systematic fashion about the relevance of particular descriptors to their learners' needs as described in Chapter 4. Rather, an attempt has been made to relate categories of description to theory, as described in Chapter 2.

In order to describe qualitative aspects of the 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) in relation to these qualitative aspects (*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 (1998: 21) this entails a certain tension between incomplete theoretical models on the one hand and operational models developed by practitioners on the other hand. The issue is complicated by the fact that, as Skehan (1995a) points out, theory has considerable difficulty accounting for the "ability for use" which is of primary interest to less specialised users. An attempt has therefore been made in this study to relate the categories of description employed on the one hand to theory, and on the other hand to the operational models.

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 a detailed survey, North (1993a) presents scales in five groups:

- brief, holistic scales of reporting overall proficiency;
- user scales reporting proficiency in different contexts of use;
- detailed, holistic rating scales;
- detailed, analytic rating scales;
- frameworks of syllabus content and assessment criteria for stages of attainment.

Brief Holistic Description

The "mother scale" the FSI could be regarded as belonging to this category because the descriptors are short and user-friendly. In North's (1993a) 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, (St John 1992; Wesche 1992) or the Finnish Nine Level Scale of Language Proficiency (Luoma 1993).

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 for the European Language Portfolio (North 1992a).

Holistic Rating Scales

Holistic rating scales 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 that the ASLPR is used to make a holistic judgement matching the performance to the most suitable descriptor, whereas FSI/ILR approach is "non-compensatory," every point in the descriptor must be fulfilled. 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.

Analytic Rating Scales

These can be seen as deriving from the "factors" considered during the FSI oral interview, and scored on a checklist 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 in early adaptations of the 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: 137), 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). This recalls Matthews (1990) complaints about feasibility and Pollitt and Murray's (1993) distinction between assessor-oriented and diagnosis-oriented scales.

Educational Framework Scales

Scales in this last 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).

Metaphors for Scales

Ingram and Wylie 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." Ingram and Wylie (1989: 2)

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 (Ingram 1985), it represents the process as a series of mastery/non-mastery decisions in relation to the thresholds between bands rather than on placing the subject on the continuum marked out by the ruler provided by the measurement scale. This draws attention to the fact 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 3.

A second metaphor has been used by Schärer (1992). 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 axes. The *scale* here is (a) the degree of reduction, also used in architectural models, scale models of cars, aircraft 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. Development can be lateral as well as vertical, language loss can be profiled as well as language gain.

> "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)

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 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." Trim (1984: 20)

Such an approach 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. These concerns are measurement issues addressed in Chapter 3.

Descriptions of Behaviour and Behavioural Objectives

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 shown 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. Behavioural objectives 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. Most criterion-referencing experts consider the effect to have been disastrous:

> "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.…. 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: 240–42)

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,* 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 strengths 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 other 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 does not mean that all scales must do so. As argued earlier a common framework is likely to be exploited in a number of different ways from a number of different perspectives and should therefore 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.

Difficulties with Common Framework Scales

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. The issues are glossed below, and then discussed in detail in Chapters 2 and 3.

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*, relateable 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 relateable 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.

Context-free : Context-relevant. A common framework entails providing a set of descriptors which are context free. 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. This is indeed the approach taken by the Association of Language Testers in Europe (ALTE). 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 (school children, 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 does not 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 Eurocentres 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. In the process comparability between sectors will be lost. It is in practice very difficult to avoid slipping into this kind of description, but by refining the descriptors with groups of teachers as described in Chapter 4 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. They may be significant functional goals in the one context and therefore practised, expected in learner performances and hence "easier." On the other hand they may not be seen as so central to another context, are not focused on and hence are considered "more difficult." 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 the descriptor 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.

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) taking account of the subjectivity in individual and group judgements (through the many-faceted Rasch model: Linacre 1989), and (d) 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, socio-linguistic competence etc. are constructs. In other words, they are creations of applied linguists which, it is claimed, have some theoretical justification." (Skehan 1984: 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 generically called 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. 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". 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 was adapted in this study as described in Chapter 4.

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 mixing methods with outcomes or by 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.

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) 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) fol-

lowing 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 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." Ellis (1989a: 310)

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. Trialling in Hawaii also 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). In addition, 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 the standards 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 standards 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, as discussed in Chapter 3. This does not 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 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.

The 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 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. A common scale seeking to provide 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. 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.

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 the scale to both descriptive theory, as discussed in Chapter 2 and to measurement theory, as discussed in Chapter 3. 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 lan-

guage as discussed in Chapter 4, and should empirically establish scale values for these descriptors in relation to the proficiency of relevant groups of learners as discussed in Chapters 5–8. 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. Chapter 9 evaluates the extent to which these requirements have been met in the current study and suggests areas for further research.

2 Description

As discussed in Chapter 1 a scale of language proficiency does not have to be holistic; many define different aspects of proficiency at different levels in what Pollitt and Murray (1993) described as a diagnosis-oriented grid. There may be occasions when one might want to give a holistic overview of a complex phenomena, as in the example given at the beginning of Chapter 1 and the arguments for and against so doing are considered at the end of this chapter. Even holistic scales, however, tend to employ categories which are reflected in the definitions for the different levels and in so doing "embody an implicit view of the construct" (McNamara 1995: 164) just as much as a series of categories presented separately. Thus whether the approach taken is holistic or analytic (Shohamy 1981), a scale cannot escape using categories which consciously or unconsciously reflect theory. The question of deciding which categories to describe is discussed in the chapter, together with the extent to which people can actually distinguish categories anyway. Towards the end of the chapter, the pros and cons of profiling and holistic rating are discussed. First, however, we consider what people mean when they talk about proficiency, and how the expressions proficiency, competence and performance relate to one another.

Definitions of Language Proficiency

Two simple intuitive definitions of proficiency are: "how successful the candidate is likely to be as a user of the language in some general sense" (Morrow 1979/81: 18), or "what is meant when we say someone is proficient in a language is that that person can do certain things in that language" (Ingram 1985: 220).

Volmer offers a more pessimistic technical definition: "Language proficiency is what language tests measure" (Volmer 1981: 152), but notes two uses of the term:

> "One pertains to *performance level* and relates then to the extent and adequacy of the learner's control of the (foreign) language in all kinds of situations and social interactions as demonstrated in tests. The other meaning unjustifiably jumps over to the *construct or competence level* of a learner or as the capacity in a given language demonstrated by an individual at a given point in time independent of a specific textbook, chapter in the book or pedagogical method" (Brière 1972: 322; Spolsky 1973: 175)." (Vollmer 1983: 5)

As Spolsky himself points out:

> "The term proficiency and the emphasis on discovering it through test performance mean that this model, while talking about knowledge, is also more likely to be oriented towards modelling the process of language use rather than towards understanding underlying competence." (Spolsky 1989: 71)

Spolsky offers a summary of the complexity of what is covered by the term "proficiency":

> "Proficiency: Knowing a language:
> Prefer to say that someone knows a second language if one or more criteria (to be specified) are met. The criteria are specifiable:
> a. as underlying knowledge or skills
> b. analysed or unanalysed
> c. implicit or explicit
> d. of individual structural items (sounds, lexical items, grammatical structures)
> e. which integrate into larger units
> f. such as functional skills
> g. for specific purposes
> h. or as overall proficiency
> i. productive or receptive
> j. with a specified degree of accuracy
> k. with a specified degree of fluency
> l. and with a specified approximation to native speaker usage
> m. of one or more specified varieties of language." (Spolsky 1989: 80)

Spolsky is here setting out choices rather than explicitly making decisions, but it is interested, for example, that he chooses to say "knowing a language" rather than "using a language" given his comment a few pages earlier (1989: 71) that the term proficiency suggests an interest in modelling

the process of language use rather than being oriented to underlying competence. In fact the only items in the list which specifically concern use rather than being an aspect of underlying competence would appear to be the word "*skill*", plus (i) *productive or receptive* and (k) *fluency*. The emphasis is also top down: measuring away from the native speaker, which is presumably why he has not felt it necessary to elaborate other qualitative criteria like complexity, appropriateness, precision; coherence etc. When discussing the use of the native speaker as criterion, it is suggested that the reason why so many scales do this may reflect the Chomskean idea of ideal competence. Yet another hint that this view of proficiency is not necessarily far from that of underlying competence is the fact that nothing in the list concerns anything outside the speaker/listener's head. A simple view of fluency is how effectively you can get whatever it is you want to express out of your head. The idea that a context, operational conditions and the creation of meaning and achievement of goals in cooperation or competition with other people are involved is lacking. In other words the dimension of putting what you have to use, the *strategic* dimension, is missing.

Interpretations of Communicative Competence

The term proficiency is preferred to competence in the context of the current study because there is such confusion over whether or not the concept of ability can be included in competence. Is competence something absolute, a property of the individual like the colour of your eyes (Taylor 1988: 153) or something relative, that you can have more or less of, can compare in a meaningful way? The confusion arises from the use of the term in two quite independent schools of thought which come together in language learning: the cognitive and the behavioural (Wiemann and Backlund 1980).

Cognitive

From a linguistic viewpoint, following Chomsky's original distinction between competence and performance (Chomsky 1965: 4), competence is seen as a mental state *excluding* ability: "it does not seem to me to be quite accurate to take "knowledge of English" to be a capacity or ability, though it enters into the capacity or ability exercised in language use" (Chomsky 1975: 23). "To know a language, I am assuming, is to be in a certain mental state" (Chomsky 1980: 48, cited in Taylor 1988: 152). Widdowson considers Chomsky "inconsistent in his use of terms by implicitly including ability in

pragmatic knowledge" (Widdowson 1989: 130), but other writers maintain he is still talking about only knowledge: "knowledge of conditions and manner of appropriate use in conformity with various purposes" (Chomsky 1980: 224) and that "for him competence is clearly a state and not a process, and has nothing to do with *capacity* or *ability*" (Taylor 1988: 151).

Many applied linguists have explicitly maintained Chomsky's distinction including Canale and Swain in their extension to include socio-linguistic competence and strategic competence (Canale & Swain 1980: 6–7) and Gumperz in his extension of competence to include socio-cultural discourse-processing conventions, knowledge of cultural norms etc. (Gumperz 1982; 1984). McNamara (1995: 163), however, perhaps following Widdowson, nevertheless puts Chomsky's pragmatic competence on the performance side rather than on the competence side, as shown in Table 2.1.

Table 2.1: Models of Knowledge, Performance and Use

Writer	Model of Knowledge	Model of Performance	Actual Use
Chomsky	competence	performance	
	grammatical competence	pragmatic competence	actual performance
Canale & Swain	communicative competence	[unable to be modelled]	communicative performance
Canale 1983	communicative competence		actual comunication
	knowledge	skill	

Behavioural

From a psychological, communication theory and educational viewpoint, it is clear that competence has been consistently taken to include both knowledge and ability: "unlike the linguistic view of competence and performance, the communication view considers performance as part of competence—not as a separate concept" (Wiemann and Backlund 1980: 188). Competence is "a combination of knowledge and skill; …proficiency in skills…is required for the manifestation of communicative competence"

normally tied to *effective* behaviour (Wiemann and Backlund 1980: 190). Dictionary definitions focus on the capacity/ability aspect (Collins 1979 edition cited by Taylor 1988: 159). The historical citations in the Shorter Oxford Dictionary for 1964 (i.e. before Chomsky) take a global view concentrating on sufficiency—of qualifications, means or income etc. Goffman relates competency in a particular instance to this idea of generalised capacity, sufficiency:

> "A competency, then, can be defined as the capacity to routinely accomplish a given complicated end. An implication is that this end could not have been achieved were the actor unable to accomplish a whole set of slightly different ones, all in the same domain of expertise." (Goffman 1981: 8 in Davies 1989: 160)

As Davies comments, the emphasis is on the skilled nature of the competencies.

In any case, Hymes extends the linguistic concept of competence in two fundamental ways (Widdowson 1989: 130): firstly by including factors other than grammar (what is formally possible) with his other three points: what is feasible, appropriate and actually performed; and secondly, even more fundamentally, by including ability for use: "Competence is understood to be dependent on two things: (tacit) knowledge and (ability for) use (Hymes 1971: 16; 1972: 282). As McNamara points out, Hymes' model is explicitly psychological and deliberately includes a range of non-cognitive factors affecting performance under his concept of communicative competence including factors "explicated in detail by Goffman"…capacities in interaction such as courage, gameness, gallantry, composure, presence of mind, dignity, stage confidence…." (Goffman 1967: 224) " McNamara (1995: 162).

Savignon, after Goffman and other communication theorists, was including *naturalness* and *poise* in assessment in 1972. She cites competency theory: study experts in order to identify "the behaviours of those considered successful at what they do, specifically, the identification of the characteristics of good communicators" (Savignon 1983: 4). Her (1972) assessment criteria do not mention grammatical competence at all, "accuracy" is message precision.

Cultural

A fact which is sometimes overlooked in representing communicative competence solely as "etiquette; speech acts; routine situational language; discourse machinery" is that although "each of these obviously forms a vital component of the concept of communicative competence, one of the most fundamental points of Hymes' argument has been ignored and that is the socio-cultural framework in which all linguistic behaviour is embedded" (Loveday 1982: 124, see also Lessard-Clouston 1992: 328). Lado's (1961: 6) diagram of the structure of language places language within culture and he speaks of grasping "from the linguistic utterance the cultural meanings encoded in it by the speaker" (Lado 1961: 5).

Widdowson's view is that socio-cultural knowledge is more complementary than central; that competence consists of schematic (socio-cultural) and systemic (linguistic) knowledge, with two forms of culturally determined schematic knowledge as highlighted by Carrell (1983, 1987): (a) *content schemata*: conceptual, ideational topic knowledge; (b) *formal schemata*: "background knowledge about the formal, rhetorical, organisational structures of different kinds of texts" (Widdowson 1983: 83–4). The more familiar the schemata associated with the content and/or mode of communication, the less reliance on systemic, linguistic knowledge. Davies comments:

> "Communicative competence is best seen as a set of scripts or schemata or ritual interchanges, plus individual differences in terms of proficiency as realised in fluency, style and creativity…what distinguishes the native speaker from the non-native speaker is what distinguished the Quaker from the non-Quaker: habit and use.…Of course there is always more to do, for native as for non-native speakers, more scripts to be acquired, more skill to be achieved." (Davies 1989: 168–69)

Parks puts the same point in a different way: it is not the *size* of one's repertoire of scripts or programming which is important but their *adequacy*: they need only be as extensive as the activities one wants to pursue (Parks 1985: 182).

Widdowson (like Gumperz) sees socio-cultural competence from a cognitive point of view, but others see it more from a behavioural perspective: Sapir's classic definition of culture is: "the socially inherited assemblage of *practices* and beliefs that determine the texture of our lives" (Sapir 1949: 207 cited in Trivedi 1978); Poyatos (1972: 64): talks of "a series of

habits shared by members of a community, learned, but biologically conditioned," Pritchard (1990: 276) of "integrated patterns of learned *behaviour*."

Competence and Proficiency

Bachman and Palmer (1984) record the resultant confusion over what people mean by "communicative competence." Davies (1989: 160) says "it is shapeless, and therefore difficult—even impossible—to define or limit for research, and especially for teaching" and that "it slides back and forth between knowledge and control (or proficiency)." Davies (1989: 167) appears to see proficiency as part of communicative competence along with innate ability and performance. He considers (1989: 169) that both the "knowledge and the control/proficiency parameters of communicative competence" are aspects of language use calling the former knowledge *what* and the latter knowledge *how*, and going on to discuss Fillmore's definition of fluency—seeing fluency as being part of the knowledge how. This is not so different from Spolsky's knowing a language, and knowing how to use a language (Spolsky 1989: 50–51). Taylor concludes:

> "…it is difficult to escape the conclusion, therefore, that when we talk about communicative competence in the context of language teaching or learning we are really talking about communicative performance." (Taylor 1988: 164)

As Parks puts it:

> "It is wise to remember that the difference between behaviour and cognition is only skin deep. Although conceptualising competence in the pure cognitive sense may have some value for a theoretical linguist, it is useless to the communication scholar who by definition must examine the world of action. …To be competent, therefore, we must not only *know* and *know how*, we must also *do* and *know that we did*." (Parks 1985: 174)

Taylor proposes the use of the term "communicative proficiency," which is reminiscent of Bachman and Savignon's proposal "communicative language proficiency" (Bachman and Savignon 1986) and goes on to define proficiency as something like "the ability to make use of competence" and performance as "what is done when proficiency is put to use" (Taylor 1988: 166).

Proficiency is here seen as something *between* competence and performance (Vollmer 1981: 160), which offers a certain parallel to Halliday's concept of meaning potential, what a speaker *can mean*, which he says is "not unlike Dell Hymes' notion "communicative competence" except that Hymes defines this in terms of "competence" in the Chomskean sense of what the speaker knows, whereas we are talking of a potential" (Halliday 1973: 54). Halliday is not actually interested in seeing language "subjectively as the ability of competence of the speaker"—which is what is of interest in constructing a scale of language proficiency. He sees the meaning potential, what the speaker can mean, being defined "objectively" by the range of options characteristic of a specific situation type (1978: 109). Nevertheless the idea of what a person can mean in a particular situation type (but might not actually perform, for various reasons) is attractive from the point of view of profiling.

As Vollmer indicated, actual usage of the term proficiency tends to overemphasise either the competence or performance aspect (Vollmer 1981: 152). One sometimes gets the impression that when the term is used in the title of a direct performance test (ACTFL *Proficiency* Guidelines/Oral *Proficiency* Interview; Australian Second Language *Proficiency* Ratings), there is a tendency to associate it more with performance. In much American literature where one talks about "the Proficiency Movement" this seems to be the case, and Ingram defines Proficiency as "not just knowledge but the ability to mobilise that knowledge in carrying out particular communication tasks in particular contexts" (Ingram 1985: 220). In Britain the eventual performance aim has always been clear: "Proficiency tests must be based upon a job analysis to determine the required behaviour" (Ingram E. 1968: 72) and there is an established tradition of direct testing, but there is not necessarily an assumption that language *for* proficient communication means exclusively testing *as* communication (see, e.g. Weir 1981: 30; Rea 1985: 26–27; Davies 1986: 61). The term "proficiency" appears in probably the most traditional EFL examination, the Cambridge "Certificate of Proficiency in English" (1913), and an Edinburgh PhD developing a Rasch model itembank to test knowledge of the language for Eurocentres had as its title: "An Item Bank for Testing English Language Proficiency" (Jones 1993). Proficiency, then, is deduced from performance, but that performance could be either of a communicative or non-communicative kind. A test of "competence" as such is difficult to visualise (Rea 1985: 18–27) if not terminologically impossible (Bachman and Palmer 1984: 34).

Bachman (1990a: 16), perhaps as a result of all this confusion, avoids using the term proficiency at all. He prefers to talk of Communicative Language Ability in order to avoid connotations of either the 1970s global proficiency concept associated with, for example, cloze tests (Carroll 1961; Spolsky 1968; Oller 1976/83), the use employed by Jones (1993) cited above, or association with the US "Proficiency movement" associated with ACTFL. His diagram of Communicative Language Ability in Communicative Language Use sees two separate knowledge bases Knowledge Structures (Knowledge of the World) and Language Competence (Knowledge of Language). These two knowledge bases are acted upon by Strategic Competence in the relevant Context of Situation to execute language as a physical phenomenon through psychophysiological processes. Thus in his view Communicative Language Use consists of Communicative Language Ability and Strategic Competence, which comes into play when that ability is put to use. Bachman's, and other writer's views of Strategic Competence are taken up later in the chapter.

Variability

Skehan (1995a) extends Bachman's argument about ability put to use and argues for the addition of another factor *Ability for Use* alongside Bachman's Strategic Competence mediating between competences and the demands of the context. This concept relates to an information processing view of language use: as attentional demands increase, speech is likely to become more pragmatic, contextual and *lexically* organised (ibid: 16)—utilising "lexicalised sentence stems" (Pawley and Sydner 1983). In other words, in a situation demanding more processing, in a effort to safeguard fluency, the learner tends to switch styles in a continuum of *learner performance* styles similar to the *learning* styles identified by Hatch (1974) and Skehan (1986; 1989a; 1991a) shown in Table 2.2.

Table 2.2: Learner and Performance Styles

rule learners	data gatherers
pattern-making problem-solvers	chunk-accumulating memorisers
(*analytic, grammatical sensibility*)	(*memory-dependent*)

Skehan's proposal is also similar to the "trade-offs" between communicative effectiveness and correctness posited by Clahsen, Meisel and

Pienemann in their Multidimensional Model (Clahsen 1980 & Meisel; Clahsen and Pienemann 1981) in which "a learner's place on the variational dimension is a consequence of how he *decides* to distribute his limited speech processing resources" (Pienemann 1987: 75, *original italics*). They identify two ends of the continuum: those who avoid deviation from target norms (Ellis 1989a/1992: error-avoiders) & those who are rhetorically expressive, citing Slobin (1975) (Ellis 1989a/1992: communicators). The Clahsen school propose that "it is possible to locate a given learner in a unique place on a continuum whose poles are the competing goals of *effectiveness* versus *correctness*" (Pienemann 1989 citing Clahsen, Meisel and Pienemann 1983). They consider that this predisposition or preference is stable so that the "learner will stick to that preference throughout a whole range of different structural problems" and that proficiency (in terms of rater impression) is explained by the interaction of the two dimensions (development stage; variational predisposition) with learners who have a "balanced style" being rated more highly (Pienemann and Johnson 1987: 90–1).

Skehan on the other hand posits a *shift in style* in relation to the demands of the communicative situation, a form of interlanguage variation dependent on context, and concludes that the concept of a unitary underlying competence cannot explain such a change in ability for use and that it is therefore "misconceived" to see competence as underlying performance in any straightforward manner. Skehan's proposal could be seen as a development of Tarone's posited spectrum of speech styles caused by the degree of attention to speech, measured quantitatively by the occurrences of certain forms. Tarone's continuum of styles is Labov's and Trudgill's socio-linguistic paradigm adapted to interlanguage (Tarone 1983), and in proposing a range from very "careful" styles to very relaxed "vernacular" styles is closely related to a prepared-unprepared continuum. Tarone and Ellis (1985a, 1985b, 1987), still working with the competence : performance dichotomy, see this kind of systematic variation as variation in competence—not in performance:

> "Planning variability is seen as a feature of the learner's competence, not just of his performance. At any one stage of development, the learner possesses two or more rules for at least some structures and calls on these differently according to whether the discourse is planned or unplanned." (Ellis 1987: 14).

Skehan (1987) considers that assessment must take this systematic variation into account. Some other linguists, however, would hold that "heterogeneous competence is simply a contradiction in terms" (Gregg 1989: 20), a position which would support those of Bachman and Skehan that there is something other than competence (in the classic meaning of underlying innate ability) which comes into play as the learner allocates and balances resources differently to meet the demands of different tasks. In this sense, Skehan's *Ability for Use* is related to Bachman's (1990a) and Faerch & Kasper's (1983) concept of Strategic Competence. The position taken by Bachman and Skehan allows for variation *within* any interlanguage style—as performance is affected by the conditions and constraints of the particular task at hand—answering a criticism of Tarone's position by Bennett and Slaughter (1985). Progress in language proficiency can be seen, as Bennett and Slaughter suggest, by both elaborating the range of styles available, and by consolidating particular styles to conform to native speaker norms as the learner becomes less linguistically distracted by conditions and constraints posed by the task at hand.

That learners *do* systematically provide different performances dependent on the task (and domain and topic of the task) is now well established. One variable is genre and text type and its relation to content domains. Swain (1993) reviews a series of Israeli studies, on reading (Shohamy 1988a); listening (Shohamy and Indbar 1991); speaking (Shohamy, Reves and Bejerano 1986) and writing (Nevo 1986), which demonstrate that performance varies with text/genre type. Such results are supported by information-processing theories of second language acquisition (O'Malley, Chamot and Walker 1987; McLaughlin and Harrison 1989) which suggest that static, declarative knowledge, including knowledge of *content schemata* (topic/domain) and *formal schemata* (discourse conventions) schemata (Carrell 1983, 1987, Carrell and Eisterhold 1983) is stored in nodes which are linked in networks to other nodes. The existence of separate networks of concepts, linkages and propositions for interactions in particular domains "is consistent with the notion of domain-specific language skills…proficiency (which) may be constructed out of experience and not established by direct transfer from L1" (O'Malley et al 1987: 291). This is related to Trim's view that a "learning biography consists not of a straight line progress from elementary to intermediate to advanced but an accumulation of life-related learning experiences." Such a variation of L2 proficiency across tasks suggests that profile reporting might be more profitable at the level of

genre/settings,"language activities," rather than at the level of the four skills (North 1992a). Naturally a move away from the four skills is made more difficult by the fact that it requires decisions about precisely which categories to replace them with, and any such choices are open to the criticism "why *this* set of functions, uses, rather than any other?" (Davies 1990: 41 discussing the ELTDU scale) without the support of considerable research to identify relevant and contrasting genres.

Another variable in relation to pedagogic tasks, is the amount and type of preparation and follow up activity. Foster and Skehan have shown that the fluency, complexity and accuracy of language produced in a task is the result of an interaction between different kinds of preparation in pre-tasks and washback (or lack of washback) from different kinds of post task with the different levels of cognitive complexity and communicative stress posed by the task (Foster and Skehan 1994; Skehan 1995b). As Skehan points out (1995b: 554) if one considers the processing implications, it is clear that it is very difficult for learners to achieve accuracy, complexity and fluency at the same time, and this leads to a need to prioritise. In relation to teaching programmes, he suggests repeated cycles of task-types with separate foci on performance (emphasis: fluency) and on development (emphasis: restructuring—i.e. adding complexity) from what might be described as polishing (emphasis: accuracy & fluency). In relation to assessment, Foster and Skehan found that one task type (personal information exchange) produces high accuracy but little complexity, whereas another (narrative) produced complexity at the cost of accuracy, whilst a third task type (decision) produced a balance between accuracy and complexity without the same occurrence of "trade-off effects" (Foster and Skehan, 1994). The very complex inter-relationships between these three task-types undertaken under different planning conditions and the degree of complexity, fluency and what they call "naturalness" of the language generated by learners who were detailed or less detailed planners (ibid 33–36), suggests that the timing of particular assessment tasks in a teaching programme as well as the choice of task and kind of planning instructions given needs very careful consideration.

Variation across tasks, in relation to preparedness and cognitive complexity is increasingly taken into account in oral testing (c.f. for example North 1986; 91; 93b as well as Shohamy, Reves and Bejerano 1986), but poses a question for reporting results. To paraphrase Alderson (1981: 61): if an individual is given various (e.g. speaking) tasks and he/she performs

them differently, does this mean that he has (a) different proficiencies which should be profiled,—a performance-related or meaning-potential related view of proficiency; or (b) a homogeneous proficiency that he/she can access to varying degrees depending on variables like familiarity with the topic, the amount of attention taken up by the processing demands of the task, the degree of preparedness etc.—a competence-related view of proficiency. Both views appear to make a lot of sense.

Conclusion

The current position in relation to the development of a model of Communicative Language Use, or in relation to the acceptance of a standard interpretation of competence, proficiency and performance remains confused. The complexity of the number of factors and variables involved in Communicative Language Use make an approach based either on a competence or a performance view difficult. As McNamara states (1995: 174) the full implications of the fact that all performance tests are taken under performance conditions subject to the kinds of issues discussed above have not either been incorporated satisfactorily into a descriptive model of language use, or taken fully into account in the design and standardisation of assessment procedures, or, one could add, considered systematically in the development of descriptors of proficiency. The lack of an adequate starting position for the development of an instrument like a scale also means that certain choices have to be made about categories, and boundaries between categories. McNamara (1995: 172) mentions the apparent overlap between illocutionary and strategic competence. Purity of categories is a problem when talking about competence/proficiency categories and about communicative activity.

The approach taken in this study and in the Council of Europe Common European Framework is to separate the consideration of Categories for Competence/Proficiency from Categories for Communicative Activities, with Strategy Use seen as a hinge linking the two. In the Council of Europe Framework, a distinction has been made between the competence categories used in the descriptive scheme in the Framework and the proficiency categories defining achievement in relation to some of those categories used in the illustrative scales of descriptors for the Common Reference Levels. The latter are derived from the study described in this book. In this study, relating proficiency categories that are meaningful for teachers and assessors to a competence model has been, to say the least, difficult. The

approach adopted takes the more behavioural view of proficiency outlined by Parks (1985), broadening the definition of Pragmatic Competence to include Skehan's *Ability for Use*, Spolsky's *Knowing how to use a language*, and Fillmore's *Fluency* cited at the beginning of this section. An attempt has also been made to take account of the issue of variability of performance conditions in the design of the rating scale used in conjunction with the descriptors.

The Native Speaker as Criterion

Before moving on to discuss categories for competence/proficiency and for communicative activity, however, since the concepts of competence and proficiency are both related to an idealised native speaker, since many scales of language proficiency take the native speaker (or educated native speaker) as a top level, and since it is difficult to talk about quality of performance without reference to some (probably native speaker) norm as a point of comparison (Davies 1989), a short digression on the suitability of the native speaker as a criterion seems appropriate.

It is presumably the Chomskean tradition, the idea of an ideal competence, which has led the majority of existing scales of language proficiency to take the native speaker (NS) as a "criterion." The NS appears in various ways. As already mentioned, the FSI scale took as its criterion the educated native speaker (ENS) or, since it became accepted that many native speakers would only reach a "3" on the 0–5 scale (Jones 1985: 82) the well educated native speaker (WENS), though Lowe attributes this development to Oscarson (1978) (Lowe 1985: 47), and judgements were made by judging the degree of nativeness or foreignness in the learners performance. The inappropriateness of using the NS (let alone WENS) from a psychometric point of view in relation to criteria for criterion-referenced assessment was pointed out long ago:

> "Users of a language differ among themselves enormously both in their implicit knowledge of language rules and in their ability to use that knowledge in language performances. In view of this, the possible existence, or even the possible theoretical description, of anything like an "ideal native speaker" is moot". (Carroll 1979: 22–23; see also Valdman 1989: 34 for a similar view)

The approach has been defended with two arguments: (a) that "most languages have a clearly identifiable group that natives, from the man in the street to a member of a group, recognise as an educated native speaker" (Lowe 1985: 47), and (b) that since some of the agencies using the ILR (Interagency Language Roundtable) may negotiate treaties, the WENS concept is necessary (Lowe 1986: 394). However, the fact that the WENS may be relevant to the US government does not mean it should necessarily be relevant to more broadly aimed assessment of adults, let alone to children.

NS competence is heterogeneous and variable

The concept of "native speakerness" itself as a target is flawed because it is not a tangible, homogeneous concept. It gets mixed up with domain variation—people are much better at some things than at others—and with language politics, prestige varieties etc. etc. Research comparing native and non-native speakers on language tests confirms this: Bachman talking of the findings of the Canadian Development of Bilingual Proficiency (DBP) project (Harley et al 1990) says: "few studies have so convincingly demonstrated the fact that native speakers vary greatly in their control of different aspects of language proficiency" (Bachman 1990b: 27).

The most obvious form of variation in native-speaker competence relates to familiarity with different "discourse domains." In an interlanguage context, a discourse domain is defined as "a personally, and internally created "slice" of one's life that has importance and over which the learner exercises content control" (Selinker and Douglas 1985 cited in Douglas and Selinker 1985: 206). As Douglas and Selinker point out, however, the concept comes from native speaker studies by, for example de Beaugrande (1984), who states that native speaker speech "on behalf of views one does not believe in has a noticeably higher proportion of errors" (de Beaugrande 1984: 28 cited in Douglas and Selinker 1985: 207).

Secondly, as Hamilton et al (1993: 1) point out, the ENS concept does not take account of the cognitive complexity of communicative tasks. Clapham and Alderson (1997 cited in Hamilton et al 1993: 7–8) demonstrate that 17 year old sixth form college students in their first year of a two year A-Level course had average scores for Reading around Level 6—the entry requirement for foreign students at British universities. In a parallel study for writing, students of about the same age on advanced vocational courses (Marketing; Secretarial) produced the same result (mean IELTS 6.5). These results are perhaps no great surprise, university lecturers often

complain about standards of undergraduate literacy (expecting the ideal, well-educated competence). But in Hamilton et al's own research, "significant differences were found between the performances of even very highly educated native speakers" whose competence appeared "far from uniform and related to educational level and work experience" (Hamilton et al 1993: 15). To be precise, of three groups, the only people who reached the IELTS "Expert User," which as they point out is a thinly disguised WENS, were junior barristers, presumably because of their training in close reading. Post graduate students and university lecturers had a better, more consistent performance than the sixth formers, but were not "Expert Users." However, one should perhaps beware of generalising too far from the results of particular tests, which may have been fallible instruments, or from test scores for a population for whom the test had not been designed. In terms of Candlin's (1987) list of factors which make up task difficulty, even operating under the assumption that the *Cognitive Load*, the *Code and Interpretative Complexity* and the *Communicative Stress* was no greater for the young native speakers involved, it might, for example, be the case that the EFL student population on whom the tests had been normed had a slight advantage in terms of recognising the *Particularity and Generalisability* of the tasks which for them may have followed a more predictable pattern as they have in general more practice at that kind of task, and in terms of *Process Continuity* an EFL test for pre-sessional students *should* fit better into the wider experience of foreign learners on EFL pre-sessional courses that into the wider experience of British adolescents.

Nevertheless, the example is salutary and highlights the point that underlying the Expert User may lurk the concept that the language varieties of more educated people are more developed, more complex—a concept disproved by any number of socio-linguistic studies. However, socio-linguistic studies tend to focus on what Cummins described as BICS (Basic Interpersonal Communication Skills) as opposed to CALP (Cognitive Academic Language Ability) in which, for example, older learners might have an advantage (Cummins 1979: 199; 1980: 179–180). Partly in response to emotional attacks on this distinction from specialists in bilingual education afraid that the "common sense" ethnocentric elitism of the ILR ENS would affect government policy on ethnic groups (e.g. Edelsky et al 1984; Martin Jones and Romaine 1986) Cummins (1983; 1984) amended his BICS/CALP distinction to a two dimensional matrix shown in Table 2.3. He admits that

these are clearly not the only dimensions necessary for a comprehensive description of proficiency.

Table 2.3: Cognitive Complexity and Contextual Support

	cognitively undemanding	**cognitively demanding**
context-embedded	engaging in a conversation	
context-reduced	writing a letter to a close friend	writing or reading an academic article

The definition of context-embedded is that "participants can actively negotiate meaning" and that of context-reduced is that one must rely "primarily…on linguistic cues and may in some cases involve suspending knowledge of the real world" (Cummins 1983: 120). Where a task is placed in the matrix depends not just on the characteristics of the task, but also on the proficiency of the user (Cummins 1983: 123). As Cummins points out, there are parallels here to Wells' study of children's development in their L1—and by extension to Skehan's study of the language aptitude of the same Bristol cohort when they encountered an L2 (French). Skehan's study showed massive differences in rate of L1 language development (Skehan 1989a: 31) and a connection between the rate of L1 language development and L2 achievement scores. This connection was linked to ability to handle decontextualised language (Skehan 1991a: 278), and to social class, parental education and vocabulary development (Skehan 1989a: 32).

One could summarise by stating that native speaker competence is not homogenous, and that both native and non-native competence vary in relation to communicative tasks (leaving aside issues like having an off-day), and hence that the level the user would be assigned on a scale of language proficiency varies in a complex manner according to the cognitive complexity and familiarity of the task and the degree of interpersonal context in relation to the age, educational experience, social class (defined by wealth and parental education) and culture/sub-culture of the language user.

NNSs are not going there anyway

The idea of "transitional competence" (Corder 1967), of interlanguage as a series of permeable intermediate stages between the first language and native-speakerness (Selinker 1972), is no longer tenable in its stronger form.

Second language acquisition studies show that learners are not on a path to native-speakerness, but rather on a path to a level of functional ability which they find satisfactory. At this point they will, consciously or subconsciously, strike a balance between on the one hand their need to do other things in their life apart from learning languages, their desire to preserve their separate cultural identity (segregative motivation: Meisel, Clahsen and Pienemann 1981: 118–9) and their personal ego-identity (e.g. *the way they sound* Guiora 1982: 173), and on the other hand their integrative or instrumental motivation towards so-called long-term accommodation (Trudgill 1981 in Beebe 1988: 66) to "in-group" socio-cultural, linguistic norms (Giles et al 1991; Ellis 1985: 256) and/or pedagogic norms (Littlewood 1981a: 154). This is not a new idea. Howatt cites Sapir that:

> "…it is quite a mistake to suppose that an English speaker's command of French or German is psychologically in the least equivalent to a Frenchman's or German's command of his native language. All that is managed in the majority of cases is a fairly adequate control of the external features of the foreign language." (Sapir 1933 cited in Howatt 1984: 7)

Howatt points out that native and non-native speakers have two quite different developmental histories comparing the learning of an L1 and an L2 to Vygotsky's (1962) relationship between inner and written speech. Learning an L2 is more like learning literacy (Howatt 1984: 8–9)—learning how to cope with cognitively more demanding, context-reduced tasks.

Lowe states that, of ILR candidates tested, less than 1% have achieved the ENS standard (Lowe 1985: 48). Coppetiers (1987: 549) reports that even non-natives like that less than 1% have a different competence to the native speakers from whom they cannot be distinguished by mistakes, range of expression or pronunciation in their performance. Scores for a group of non-native speaking professional linguists living in France on a grammatical intuition task (a sort of true/false task on sentence acceptability focusing on problematic linguistic contrasts like passé simple, passé composé) did not overlap at all with the score range for a group of native speakers (their husbands/wives and native speaker professional associates). The two data sets showed distinct qualitative and quantitative differences in grammatical intuitions. Although some SLA researchers express reservations about grammaticality judgements (e.g. Ellis 1991a) this reflects Howatt's point. Coppetier (1987: 569) explains the difference by reference to different development histories: native speakers as children store instances of words

as separate, very redundant entries and only gradually (and still subconsciously) replace this with a more analytical system, perhaps retaining reflections of the contextually rich environment in which the forms were initially learned. This view is supported by research on French schoolchildren who apparently "accumulate a storehouse of related forms in memory, but never make the inductive leap to abstract rules (*about gender*), either consciously or unconsciously" (Schmidt 1990: 146–8). The result is a comprehensive network which acts *as if* it knew the rules ("parallel distributed processing": Rummelhart and McClelland 1986) a network which L2 learners cannot aspire to, even though it may be the case that they develop their own more limited networks of *exemplars:* "specific contextually coded items which may contain structure, but which are learnt as chunks" (Skehan 1995b: 547 citing Carr and Curran 1994).

There *is* a counter argument, however. Davies considers the native speaker "a concept which is, in spite of its fuzziness, essential" (1989: 158). In discussing meanings of the phrase "language ability" he recognises that we consider some native speakers to be more fluent, better speakers —one way in which we use the term—but that when we talk about second language learners, "language ability" always implies some comparison to native speakers (goal) in deciding an acceptable level of performance at a particular level (standard)—that both aspects are essential to criterion-referenced assessment and that "in both cases, in the goal and the accepted level, there is substantial uncertainty" (Davies 1990: 52). "The native speaker is not made any less necessary by being elusive, nor is the concept made any less elusive by being necessary" (1989: 159). In this view, any judgements of accuracy, socio-linguistic appropriacy, socio-cultural savoir-être, discourse conventions etc. etc. can only be made by reference to the norms of the native speaker culture(s), the implicit or explicit goal of language learning.

A tentative conclusion to the controversy might therefore be that while it is not only unnecessary but in fact misleading for scales of language proficiency for foreign language learners to have a top level of the "native speaker" or "expert user" in defining standards for degrees of skill of performance at different levels of proficiency, reference will have to be made either to native speaker norms or to interlanguage norms. Since the latter approach would carry the dangers of (a) inviting underachievement (since studies have shown that above all children are very influenced by what their teachers expect of them, a basic problem with all minimal competence standards), and (b) increasing culture shock, misunderstanding and performance

anxiety when native speakers are met, thus undermining any feeling of success, such a radical step appears unjustified. Therefore, despite the elusiveness, as Davies says, as a point of reference for defining what is acceptable at different levels, there appears to be no alternative to the standard native speaker variety. Poyatos (1972: 74) defines the standard socio-culturally as what is "common to the refined and the rustic, to the educated and the pseudo-educated" distinguishing this from what is particular to the repertoires of particular discourse communities within the speech community and from the infrastandard: what would be unacceptable to most people under normal circumstances (bad language).

One is, nevertheless left with the problem: Which native speaker—and who counts as a native speaker? In a diglossic context like the German-speaking parts of Switzerland this issue can become very sensitive in relation to German; there is a variety of High German, Swiss *Schriftdeutsch*, which is not identical to *Bundesdeutsch*. For English there is the further complication: "Whose English?" with the related problems "What culture? Which culture?" (Prodromou 1992) in terms of the choice between US and UK English and the use of English as a lingua franca, even within Switzerland. However, Davies' argument still holds for the use of English as a lingua franca: when third parties use a language as a lingua franca, some (native speaker) point of reference is still necessary to ensure mutual intelligibility. The fact that for one person that reference point may be British English (maybe twice removed) whilst for the other it is American English (from a stay abroad) might be posited to cause some problems, but in a Swiss educational context, it is the British variety which is in practice privileged.

User Perspectives

Any attempt to describe competence/proficiency at a number of levels must address the question of what categories to use, and as Alderson (1991a) has pointed out with his distinction between constructor-oriented, user-oriented and assessor-oriented scales, it is unwise to divorce the development of categories from the question of how the system developed is likely to be used. Clark (1987), speaking from a teacher training perspective, summaries the choice of categories for profiling general language achievement as follows:

- tasks or macro-functions e.g. transacting, establishing and maintaining a relationship, discussing, narrating, searching for information, corresponding etc.;
- the four skills;
- hypothetical constructs of communicative competence: e.g. linguistic competence, socio-linguistic competence; discourse competence; strategic competence (Canale and Swain 1980)—to which one could add more observable or at least teacher-friendly aspects of proficiency like range, accuracy, pronunciation, fluency;
- some sort of pragmatic mixture of various sorts of categories derived from the instruction and communicative syllabus.

Systems choose between these options according to their context and pedagogic culture, aims and public. As Clark goes on to say the resultant decisions are largely pragmatic:

> "The categories we create for descriptive and referential purposes are so convenient and so indispensable that the temptation to accord them ontological status is great. But they are nothing but conceptual artefacts" (Monippally 1983 cited in Clark 1987: 40).

Summarising Clark's list there are two sides to the issue: on the one hand the (*constructor/user oriented*) descriptions of proficiency in different contexts of use (tasks/macro-functions/skills), and on the other hand the (*assessor/diagnostic-orientated*) categories of the analytic qualities or aspects of that proficiency one wishes to focus on. In Eurocentres, a simple distinction is made between *insider* (assessor/diagnostic) scales which focus on aspects of quality and *outsider* (user) scales for certification. McNamara (1990, cited in Elder 1993: 248–9) provides empirical evidence for this posited difference of perspective: non-language specialists (*outsiders*) focus on what he calls a "strong" approach in terms of successful task completion, whereas teachers (*insiders*) go more for a "weak" approach: the quality of the language sample.

This finding is consistent with research on the reactions of native speakers to learner performance. There are indications that naive native speakers in practice pursue a holistic strategy in judging competence, being most influenced by an impression of fluency (Lennon 1990) and overall intelligibility (Ludwig 1982 cited in Brindley 1989: 122), in relation to the task in hand (Brindley 1989: 123), by appropriate, or at least non-offensive,

socio-cultural behaviour (Oksaar 1992: 15), by an ability to keep going, focused on the communication rather than distracting with lengthy pauses and clumsy communication strategies (DeKeyser 1989: 115)—what Goffman (1955/72: 322) and Savignon (1972) have called "poise". Overall it appears that (a) "naive (native speaker) raters are not capable of separating one aspect of non-native discourse from another when they are asked to specifically evaluate just one" (Varonis and Gass 1982: 131) and (b) naive native speakers of most languages appear significantly more tolerant of grammatical errors than teachers, especially non-native teachers (Ervin 1977 cited in Eisenstein 1983 (Russian); Galloway 1980 (Spanish) and Pollitzer 1978 (German) cited in Loveday 1982: 147; Hughes and Lascatatou 1982 cited in Clark 1987 (English: Greek EFL).

When one considers both McNamara's (1990) findings about *outsiders* and the research referred to above, it is not surprising that scales written for carrying out language audits and orienting training programmes in business contexts have long focused purely on task completion, as both the IBM France charts in Trim (1978) and the ELTDU scale (ELTDU 1976) suggest.

Therefore, in order to provide transparent reporting of achievement, a common framework scale should incorporate descriptors not just for those aspects of competence/proficiency which are of interest to *insiders* (*diagnosis-oriented*: Pollitt and Murray 1993), but also descriptors for task completion. The latter could be used by *insiders* as a source of content for syllabus organisation and continuous assessment (*constructor-oriented*: Alderson 1991) as well as, probably in a summarised form, for reporting results to outsiders (*user-oriented*: Alderson 1991). Both types of information can be relevant to learners for orientation and/or self-assessment. If motivated learners are brought *inside* the development process through language awareness activities and self-assessment procedures to monitor the quality of their performances, then *insider* descriptors on aspects of those performances might become as relevant to them as to their teachers.

This implies the existence of a set of categories to describe communicative competence/proficiency and the existence of a set of categories to describe the communicative activity which is the context of language use.

Categories for Communicative Language Proficiency

As discussed earlier in the chapter, a satisfactory model of communicative language use has not yet been established. The basic problem is that the Chomskean distinction between competence and performance has difficulty

in coping with what happens when competence is put to use, and is in any case only one interpretation of the word competence. Variability due to performance conditions cannot be relegated to the vagrancies of imperfect actual performance because ability for use in communicative acts is (a) connected to non-cognitive abilities and skills, and (b) the systematic result of trade-offs between different processing demands caused by those conditions. The mobilisation of competences and skills in order to achieve a certain goal within the processing restraints of a communicative task has been described as strategic *competence* (Bachman 1990). Yet such mobilisation clearly does not fit within a Chomskean competence / performance model since it can only be realised in language use. The confusion over the area between competence and performance is particularly significant for pragmatic competence, which Chomsky added later to his model. Some applied linguists (e.g. Taylor 1988) assign this definitely to "knowledge"—i.e. the competence side of the competence / performance dichotomy, whilst others (e.g. McNamara 1995) consider it a middle category, the precursor of a model of performance rather than part of fallible, actual performance.

In this section these issues are explored again from a different perspective: the aim of deriving a set of categories for aspects of proficiency which can be related to a competence model and yet be used to describe language use. First the major competence models are compared. This is then followed by a discussion of the problem of overlap between categories, and then finally major categories are considered in turn.

Canale and Swain (1980, 1981) developed their model of communicative competence as an extension of the Chomskean tradition, and are thus in the same school as Gregg in believing in a competence which is homogeneous, i.e. one which does not vary in context, though the performance based on it does. However, this homogeneous competence is made up of different components—or aspects. With regard to components, there is a considerable amount of overlap between what might be considered the three leading theoretical models, those of Canale and Swain, (modified by Canale 1983), and those of Van Ek (Van Ek 1986, Van Ek & Trim 1990) and of Bachman (Bachman and Palmer 1982, Bachman 1987, 1990a) which appear related to it. In all three models, the delineation between pragmatic, discourse and socio-linguistic competence is not always clear, and all three groups of authors have shuffled the grouping of categories in succeeding versions of their models.

The Bachman model differs from the other two in that it is a "working model" designed to show how tests operate in practice. Communicative Language Ability is only one aspect of the Bachman model. It is in fact more noted for establishing that other factors and in particular test method facets radically effect test performance, and for the centrality which it gives to strategic competence, which is seen as *separate* from language competence. The three models might be related to one another as shown in Table 2.4.

Table 2.4: Models of Communicative Competence And Language Ability

Canale	**Van Ek**	**Bachman**
Grammatical competence: (lexical items, rules of word formation, sentence formation, literal meaning, pronunciation and spelling)	Linguistic competence 1. Language functions 2. General notions 3. Specific notions 4. Grammar & intonation 5. Vocabulary and idiom	A. Language knowledge - Illocutionary comp. - Grammatical comp. (Lexis, morph., syntax)
Socio-linguistic competence (appropriateness of meanings and forms)	6. Socio-linguistic competence	- Socio-linguistic competence
Discourse competence (cohesion and coherence)	7. Discourse competence	- Textual competence
Strategic competence (enhances the rhetoric effect of utterances)	8. Compensatory competence	B. Strategic comp. (Assessment, Planning, Execution)
	9. Socio-cultural competence	
		C. Psycho physiological Mechanisms: (Mode: recep, prod.; Channel: oral/aural, visual)

Sources: Canale (1983: 339); Van Ek & Trim (1990); North et al (1992); Bachman (1990a; 1991)

As can be seen by the layout and content there is a very considerable degree of agreement, but also some substantial differences. In terms of layout, Bachman separates Strategic Competence completely from Language Knowledge (henceforth referred to as Linguistic Competence) and takes a far broader view than Canale or Van Ek. Van Ek and Bachman are in agreement in placing structural, functional and lexical language knowledge in Linguistic Competence, and broad agreement on Socio-linguistic Competence can be deduced from the way it is glossed in the Canale model. All three place Discourse/ Textual Competence in a similar fashion, though Van Ek takes a broader view and includes discourse functions which could be called conversational management strategies here. Finally Van Ek is the only one to separate Socio-cultural from Socio-linguistic Competence.

As McNamara (1995: 172) notes, Bachman's model makes a significant step in that "a start has been made on acknowledging the role of aspects of "ability for use" in performance, though in Bachman (1990) these are restricted mainly to general cognitive factors.

Attempts at Empirical Validation

Attempts to confirm the supposed structure and components of proficiency posited by such models by operationalising them in tests have been very limited in their success. Such empirical studies initially used different forms of *factor* analysis, Oller's studies being the most famous (Oller 1976/1983a; Oller and Kahn 1981). Then as the limitations of these became clearer (e.g. the fact that exploratory factor analysis will virtually always create a large first factor, see Carroll J.B 1980: 588; Vollmer 1981: 170; Vollmer and Sang 1983: 68) a multi-trait multi-method (MTMM) methodology based on confirmatory factor analysis tended to be adopted (Bachman and Palmer 1981, 1982).

Bachman and Palmer's original multitrait-multimethod study of grammatical competence, pragmatic competence and socio-linguistic competence (the traits) through a modified ACTFL interview, writing sample, a multiple choice test and a self rating (the methods) found a higher order general factor plus two trait factors which they called grammatical and pragmatic competence (Bachman and Palmer 1982). The Development of Bilingual Proficiency project at Toronto (Allen et al 1983; Harley et al 1990) tried to take things a step further by again using confirmatory factor analysis in an attempt to empirically validate the Canale and Swain Model (Canale 1983 version): grammatical, discourse, and socio-linguistic competence (pre-

sumably seeing strategic competence as "compensatory" like Van Ek), plus Cummins BICS / CALP and Bialystok's (1978, 81, 82) distinction implicit learning / explicit learning. The results were disappointing, failing to support the hypotheses.

Bachman attributes the failing to the test construction, suggesting that the instruments were more complex than the traits they were trying to measure—that the model got mixed up in its operationalisation. Secondly he criticised the use of a rotation approach, orthogonal rotation (normally used when traits are not expected to correlate). Thirdly he suggested that the interference of test method probably accounted for as much variance as the traits being measured, even if they had not got mixed up anyway (Bachman 1990b). Schachter (1990: 44) at the same Symposium, argues on theoretical grounds for a basic grammatical/pragmatic distinction, and also argues that the model was not conceptually clear and therefore imperfectly operationalised, hence the finding of one large factor. Attributing the large factor to a general proficiency as Oller had done (Oller 1976) was to miss the point: the problem is in the conceptualisation (Schachter 1990). Paulston (1990) is a little more blunt than either Bachman or Schachter:

> "Another research issue is the law of the hammer. Give a small boy a hammer and everything he encounters needs hammering. In the DBP model validation studies the hammer was factor analysis, and I found it interesting how very little elucidation results from the analysis. Any study that can have three mutually exclusive "solutions" leaves me confused.
>
> > "An inherent difficulty in validating models of L2 proficiency is that measures faithfully reflecting a particular construct may not have adequate psychometric properties, while other psychometrically acceptable measures may fall short of representing the construct." (Harley et al 1990: 24)
>
> The implication is quite clear that we need qualitative and quantitative approaches to understanding second language acquisition; and that any reliance on quantification and psychometrics, however rigorous, is not sufficient." (Paulston 1990: 190–1)

Other studies, concerned with the direct rating of proficiency, also suggest that attempts to trap underlying parameters of language ability through quantitative methods like factor analysis and multitrait-multimethod analyses have only limited chances of success. For example, two studies using factor analysis of performance ratings from behavioural scales of work performance (Norman and Goldberg 1966 and Kavanagh et al 1971) suggested

that "a factor analysis of ratings tells us more about the cognitive structure of the raters than the behaviour patterns of the ratees" (Landy and Farr, 1983: 155). In other words, the parameters people see may reflect more the way they think than the parameters in what they are looking at. This reinforces findings from two studies on "halo effect" (transfer of judgement from overall holistic rating to rating for specific categories, or between categories) which suggested that "observed attribute intercorrelations may be at least partially a product of raters' conceptual schemes as well as of the true intercorrelation between traits" (Cooper 1981: 223 summarising Passini and Norman 1966, and Norman 1963). Borman has therefore argued that it is a mistake to use different partners' ratings as the methods in a classic multitrait multimethod analysis to establish the validity of performance ratings, expecting them to display convergent validity (i.e. agree on ratings), since in work performance the partners (supervisors, peers, subordinates) have legitimate separate perspectives, and may well be concerned with different aspects of performance and therefore "may capture different aspects of the total criterion construct space" (Lance et al 1992: 447).

Strong halo effects are also shown in a study (Hamp-Lyons and Henning 1991) attempting to use multitrait multimethod analysis to validate definitions of rating qualities designed to give communicative writing profiles (with qualities as traits and raters as methods). Rater judgements on the different qualities were related to one another, and did not confirm the independent existence of the qualities.

A study by Pollitt and Hutchinson describes another occasion on which raters could not rate qualities separately. Their study used a three component approach to the assessment of writing skills which focused on Appropriacy (equated with socio-cultural competence); Ideas Structuring and Selection (equated to discourse competence), and Expression (equated to grammatical competence), with the socio-linguistic element determined by the context, audience and purpose of the tasks: a letter, a report, a newspaper article, a story, or an opinion (Pollitt and Hutchinson 1987). Two forms of analysis were used: firstly a traditional correlational analysis and secondly a Rasch analysis. The correlations showed that the ratings for the underlying competences (performance qualities) were strongly inter-related *and* that performance on one task appeared to be almost completely independent of performance on another. They conclude that:

> "The results underline the importance of including a wide enough range of language functions in writing tasks in any comprehensive language assess-

> ment, while at the same time they suggest that the particular model of competence used *(to rate the performance on different subscales)* may not be too important." (Pollitt and Hutchinson 1987: 90)

They also discuss in some detail the way the Rasch model yielded far more interesting information than correlations about the way in which the performance level, the tasks and the competence components (performance qualities) interacted in patterned ways. In other words, although the Rasch model is concerned with measurement, it is in practice *more informative* about the structure of competence at different levels than traditional methods which reduce everything to numbers like correlations, multitrait-multimethod analysis and factor analysis.

The fact that all these studies do not empirically demonstrate a consistent structure of competence / proficiency is at least in part to a failure to distinguish adequately between *components* of competence, which exist separately, and *aspects* of competence (Shaw 1992: 10) or *areas* of knowledge (Bachman and Palmer 1984: 35), which do not necessarily do so. Van Ek maintains this distinction in talking of socio-linguistic, discourse, socio-cultural competence etc. as *parameters* of description, *aspects* of competence, not components (in North et al 1992: 17), though earlier he also tended to use the term component (1986: 33-37). The Canale and Swain model had previously been criticised in this regard:

> "It has been suggested that communicative competence includes linguistic competence in the Chomskean sense (e.g. Canale and Swain 1980.) This, I think, is misleading if the latter is conceived of as in some sense a separate *component* within the former." (Widdowson 1984: 102)

The DBP project members implicitly accept this point and now talk of aspects like socio-linguistic competence as heterogeneous traits which consist of a number of sub-components which may not correlate (Harley et al 1990: 51). Swain (one of the DBP team), now considers that the team:

> "…were seduced by testing theory into believing that we should be finding high internal consistencies…. *whereas* low levels of internal consistency are indicative of the complexity of second language proficiency." (Swain 1993: 196)

The problems of not distinguishing between components and aspects is mirrored in the failure to distinguish between components (or maybe one

should now say aspects) and *factors* (Cziko 1984: 24). The Canale and Swain or Van Ek models are "descriptive models" whereas Oller's model (Oller 1976/1983a) and Bachman's model (Bachman 1990a: 82–106) are "working models" which attempt to show how components of communicative competence are interrelated psychologically to form a set of independent *factors*. The problem with working models is that they "come and go" (Skehan 1991b: 15) and, as stated in the introduction to this section, we are still a long way away from a model which explains communicative language use.

MTMM analysis has been the main methodology for these attempts to explore empirically the structure of proficiency or competence and might therefore be thought to be relevant to the empirical development of a scale. But as Carroll points out, the fundamental flaw of MTMM studies like those of Bachman and Palmer is that they are concerned "not so much concerned with the differentiation of language skills as with the construct validity of the measurements that are developed" (Carroll J.B 1983: 96). In fact, MTMM seems to have little to offer in relation to the development of framework descriptors for several reasons.

Firstly it is very difficult to operationalise the theoretical constructs in tests which keep them separate. They appear not to be homogeneous traits,—as the DBP project discovered. D'Anglejan remarks that Halliday's work suggests a big overlap between elements of competence, and that they "cannot be operationalised, other than trivially, in ways that make them amenable to empirical validation" (D'Anglejan 1990: 148). In such a situation, an MTMM analysis is not helpful: "If items from the same scale actually reflect different traits, or items from different scales actually reflect the same trait, then scale scores cannot be interpreted in terms of trait and method effects" (Marsh and Hocevar 1988: 108, cited in Lance et al 1992: 439).

Secondly, it is very difficult to stop the method effect of the type of test or rating system influencing the results. Davies commented early in the process: "No method it seems to me can ever be entirely free from the trait it seeks to realise" (Davies 1981: 184). This helps to explain why MTMM work has been "less successful than hoped" (Spolsky 1990: 10).

Thirdly, it is very difficult for raters with the same perspective to keep constructs separate when rating performance because of (a) the fact that the analysis may show what was in raters' heads more than what was in the performances; (b) so-called "halo effect" holistic rating; (c) the fact the traits may be non-homogeneous and therefore genuinely interrelated—so-called

"true halo;" (d) the fact that the student performances on a scale of proficiency may very well yield predominantly flat profiles since the steps on most scales are relatively large in relation to learning development; (e) cognitive overload (Matthews 1990a).

And finally, the results obtained from traditional quantitative methods like factor analysis and MTMM are in any case dependent in a variety of ways on the sample of learners used (Farhady 1982: 55; Carroll J.B. 1983: 93; Upshur and Homburg 1983: 194; Cziko 1984: 28 & 34; Sang et al 1986: 60 & 70) and the way in which the teaching matches their learning experience and style (Sang et al 1986). Unlike Rasch these classical statistical techniques do not offer objective, sample-free, instrument-free measurement (Wright and Linacre 1987: 2).

Bachman's overall conclusion is that, in the terms of the outline of the history of empirical research into the nature of language proficiency, in the same way that Oller's research on a global "g" factor (Oller 1976) and his retraction (Oller 1983) brought to an end the era of exploratory factor analysis, the DBP study brings to an end a second period characterised by "increasingly complex and comprehensive frameworks of language proficiency" and by "sophisticated (and exhausting) research designs and statistical analyses." He continues:

> "During that time (the second period) several other studies have also demonstrated what, it seems to me, is one of the main outcomes of this study: that both the background characteristics of language learners and test method effects can influence test performance as strongly as the traits we wish to examine, and that there are thus limitations on the analysis of test performance as a paradigm for research into the nature of language proficiency. While there may still be a researcher or two out there who secretly hopes for the opportunity to conduct a "really big" MTMM (multi-trait-multimethod) study, the DBP MTMM study may have marked the passing of a paradigm…." (Bachman 1990b: 37–38; 1990a: 353–4)

The result is that one can expect research to concentrate on interpreting language performance "explicitly in terms of how different aspects of communicative competence have developed as a function of specific language acquisition/learning experiences" (Bachman 1990b: 38; 1990a: 354) rather than trying to make general claims about the structure of proficiency—or competence. Multivariate statistics which permit casual modelling e.g. the LISREL program as employed by Gardner's team in their research into

learner characteristics, especially motivation (Gardner 1985; 1988; 1991; Clément et al 1980) can be expected to provide insights into the relationship between learner characteristics, second language acquisition and the shape of the resulting proficiency. But, as Gardner admits in relation to his socio-educational model of second language acquisition, the results can be expected to at best partial, and as has been pointed out in relation to his findings on integrative and instrumental motivation, limited to the particular social milieu in question (Ellis 1985: 118; Beebe 1988: 70; Skehan 1991a: 283–5) and not necessarily immune to the problems of method effect (Bachman 1989a: 201).

To paraphrase, any statements about the nature or structure of proficiency will apply to the population of learners involved; thus validity will be seen as relative (Henning 1990: 379).

Intersubjective Mapping

To put this another way, the validity of any framework to describe proficiency which is based upon the analysis of data from tests is bedevilled by (a) the problems of isolating and operationalising the desired construct in a test item, as the DBM project discovered, and by (b) the problem that the data will in any case reflect the characteristics of the particular learner population. If assessment is by subjective rating rather than tests, then any framework to describe proficiency which is based upon the analysis of data from subjective rating is bedevilled by (a) the problem of whether people really can isolate particular constructs, which is discussed in relation to "global language proficiency" and (b) the problem that the data will in any case reflect not only the characteristics of the particular learner population, but also the concept of the construct, the perspective of the population of raters.

If a shared rater perspective is exploited in order to create a defined frame of reference, as in this study, then again, statements about the validity of the descriptive approach used will need the proviso "when used by the type(s) of raters concerned." If the framework of reference is objectively scaled with a measurement model, as in this study, then that is an objective scaling of a subjective consensus. It does not necessarily reflect the nature or structure of proficiency, even of the group of learners involved. The most that can be said is that it maps that proficiency with objective scale values in terms of categories, perceptions and conventions which are shared by the group of raters concerned.

It can be argued that the nature and structure of proficiency is in any case so complex that all the various factors can only be balanced in an informed subjective judgement. This is the position held by Ingram, who claims that such subjectivity can be held to be necessary because of the complexity and redundancy of language and the need to compensate between strands of competence developing at different rates (Ingram 1985: 222) and, most of all, because language is not the sum of its parts (ibid: 233).

Ingram's position has been attacked by the proponents of the Multidimensional Model of SLA because they interpret his statements that proficiency scales like the ASLPR "seek to define proficiency levels by describing the language behaviour observable at different stages as learners develop from zero proficiency to native like" (Ingram 1985: 221) with proficiency "viewed developmentally (a diachronic, psycholinguistic perspective)" (ibid: 230) as a claim that the ASLPR can measure language development, i.e. second language acquisition. They criticise proficiency scales (Pienemann and Johnson 1987: 91) and the tests used in the Canadian DBP project (ibid: 97) as "rubber rulers." They describe the components written into descriptions on proficiency scales as "vague, intuitive, and, most importantly, relational (e.g. effect on listener)" and they go on to conclude that:

> "While proficiency continues to be defined in such terms, assessment of communicative competence can only be properly interpreted as a mapping of behaviours—that of testers on the one hand and testees on the other" and that inter-rater reliability statistics are no guide to the validity of the construct of proficiency involved, but "establish only a degree of consistency in the behaviour mapping referred to above." (Pienemann and Johnson 1987: 67)

The fact that rater perspectives are incontrovertibly entwined with ratings of proficiency and that categories are to that extent "intuitive" and "relational" has been discussed above. Fluency is a good example of a category of this type. Lennon has defined Fluency as:

> "…an impression on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and efficiently.…To some extent then, fluency reflects the speaker's ability to focus on the listener's attention on his or her message by presenting a fin-

> ished product rather than inviting the listener to focus on the workings of the production mechanisms." (Lennon 1990: 392)

Fulcher (1993) has demonstrated that a Fluency rating scale based upon features occurring in the discourse of the type of students concerned is an effective predictor of proficiency, and as will be reported later, despite some difficulties in workshops with teachers in deciding what Fluency is, Fluency turned out to be the category which was interpreted most consistently and predictably across educational sectors and language regions in the questionnaire survey of learner proficiency. This is, in retrospect, not so surprising considering the holistic nature of Fluency.

To summarise: in making judgements of proficiency, an expert—and raters are experts—synthesizes the various strands of competence he/she observes into a holistic judgement, taking account of the two dimensions represented in the SLA Multidimensional Model: [(i) stage of development; (ii) trade offs between effectiveness and correctness] which might be glossed as fluency and accuracy in teaching terminology (Brumfit 1984). "Experts" in a field are known to synthesize complex multiplicity into higher order categories so as to be able to take strategic decisions (McLaughlin, Rossman and McLeod 1983: 136–7; Nation and McLaughlin 1986: 42). What is being observed is clearly complex multidimensionality but experts synthesize, and one can argue that such synthetic subjective judgements are essential to take account of the complexity of the phenomena of language (c.f. Read 1981: x; Ingram 1985: 252). The difficulty is that it is very difficult to establish exactly how experts synthesize, and it is relatively clear that they do it in many different ways and can therefore arrive at different results (Einhorn 1974; Meyer and Booker 1991). Hence concerns over inter-rater reliability with subjective approaches to assessment. But if a way can be found to identify and quantify the consensus values given to different criteria, to different descriptors, and to take intra-rater consistency and inter-rater variability in severity into account in assessing learners stage of attainment, then a calibration of an inter-subjective map can produce a defined common framework of reference: common, that is, to the raters who produced it, and to raters like them, in relation to the learners who were assessed, and to learners like them.

Categories used in the Study for Communicative Competence

The competence model used as a point of departure for the study, and for the Council of Europe Common European Framework, was the Canale and Swain, Van Ek, Bachman model with four main aspects: Strategic Competence; Linguistic Competence; Pragmatic Competence, and Socio-cultural Competence. The rest of this section considers each of these competences in turn, and then discusses the issue of accommodation (speech adjustment), which is negatively correlated to proficiency. The approaches of the three models compared are discussed below within the context of the literature on communication strategies and a set of categories for descriptors of Strategic Competence is put forward. This is followed by a discussion of various common pitfalls in the description of Linguistic Competence, and a discussion of the complex issue of Pragmatic Competence. It is suggested that in the absence of a fully fledged model of communicative language use, "ability for use" and fluency might best be situated here in relation to competence model. Finally difficulties of defining and scaling Socio-cultural Competence are described.

Strategic Competence. Since Canale and Swain's model it has been generally accepted, at least in theory, that one needs to take into account "the communicative strategies that are developed in communication and which constitute communication ability" (Holec 1980: 30). As seen in the comparison of models of competence, Bachman gives strategic competence a central position as "a general ability, which enables an individual to make the most effective use of available abilities in carrying out a given task, whether that task be related to communicative language use or to non-verbal tasks" (Bachman 1990a: 102; 106). Bachman relates his concept of strategies to that of Faerch and Kasper, who consider that: "If one accepts the basic distinction…between the planning and the execution of speech, communication strategies can best be placed within the planning phase, more precisely, within the area of the planning process and the resulting plan" (Faerch and Kasper 1983: 30). Faerch and Kasper's own definition reads: "potentially conscious plans for solving what to an individual presents itself as a problem in reaching a particular communicative goal" (Faerch and Kasper 1983 cited in Bialystok 1984). Defining, describing or finding ways to assess that strategic competence is not a simple matter, however. Faerch has noted: "There is considerable disagreement as to whether strategies should be considered a particular type of psycholinguistic

process (Selinker 1972) a particular type of psycholinguistic plan (Faerch and Kasper 1983) or a particular type of interactional process (Tarone 1981/83)" (Faerch 1984: 50), and the plethora of taxonomies and lack of clear distinctions between learning strategies (learning to learn) and communication strategies (an aspect of proficiency) have not helped matters. Furthermore, as the Canale and Swain and Van Ek models suggest, there was a tendency in earlier work on strategies to focus narrowly on what have been called communication compensation strategies: "the learner's attempts at bridging the gap between…resources and the message to be conveyed" (DeKeyser 1989: 108), though early workers in the field later widened their definition of strategic competence (Tarone 1983: 64; Canale 1985: 5).

Perhaps as a result of all this confusion, of the 41 scales of language proficiency for spoken interaction included in North's (1993a) survey, of which 27 were developed after Canale and Swain's model became available, only 3 take strategies as a category (Gothenburg; RSA Modern Languages 1989; Cambridge Assessment of Spoken English (Milanovic et al 1992/6). In one of those (CASE) strategies are a sub-category of interaction. So from 27 post Canale & Swain scales, only 2 treat strategies explicitly, and both confine themselves to repair or compensatory strategies.

This is perhaps not that surprising since strategic competence remains such an elusive concept, with the chicken and egg question of whether people with "good" strategies advance, or whether advanced people use "good" strategies, essentially unanswered. Even though several studies have found that more advanced students do use more L2 based strategies and/or metacognitive strategies (e.g. Faerch and Kasper 1983: 76; O'Malley et al 1985: 37; North 1986; Fulcher 1988; Rost and Ross 1991: 242), and while Willems talks of a hierarchy showing "growing complexity of a verbal nature" from primitive paralinguistic, through L1 based to interactive L2-based strategies like checking questions and initiating repair (Willems 1987: 356)—a hierarchy which Chesterfield and Chesterfield (1985: 54) even present after a cross sectional study as an implicational scale—it is at the same time clear that different task conditions cause wide variation in strategy use (O'Malley et al 1985: 40; Poulisse 1987 cited in Lennon 1990: 397).

Talking about the relationship between *learning* strategies rather than communication strategies (though the distinction is, like many distinctions not watertight) Ehrman and Oxford (1990: 312) conclude: "Research indicates that language learners at all levels of proficiency use strategies, but some are relatively unaware of those they use. More proficient learners ap-

pear to use a wider range in a greater number of situations than do less proficient learners, but the relationship between strategy use and proficiency is complex." The implications is that the difference between learners who learn well and those who do not may be not so much the strategies themselves as "the flexibility and appropriateness with which strategies are used" (Chamot and Kupper 1989 cited in Skehan 1991a: 288). Again this comment is made in relation to learning, rather than communication strategies but might apply to them as well. As the following discussion suggests, it is perhaps not so much *what* you do as *when* you select it as an appropriate course of action—which comes back to the issue of planning.

Unfortunately, strategies cannot be observed, only the partial traces left by *some* strategies can be discerned in speech, data or students' memories. Since learners, especially advanced learners, try and predict communicative problems at planning points in their discourse and either steer round them or adopt a *formal reduction strategy* (Stick to simple means to avoid breakdown, but try and say what you want to) or a *functional reduction strategy* (Compromise your message, reduce it to something you can express), it is very difficult to get accurate data on communication compensation strategy use without using self-report data (Faerch and Kasper 1983: 198). By the same token, it is then difficult to observe and evaluate strategy use and self report data is notoriously unreliable when there are stakes involved, as there would be in assessing a band on a scale. Reduction strategies are observable when a search for fluency results in simplified interlanguage forms, (full formal reduction) and there has long been plenty of anecdotal evidence suggesting that this happens. Coulter talked of two elderly Russians who avoided articles, plurals and past tenses because they knew from experience that if they thought about grammar while speaking, they would appear hesitant, and native speakers would get impatient with them (Coulter 1968 in Selinker 1972 in Richards in Pride 1979). More recently DeKeyser compared native speaker reactions to two learners with a similar knowledge of grammar and vocabulary:

> "The learner who was sought out for conversation by the natives camouflaged his constant monitoring, used inconspicuous communication strategies and filled pauses. (…) The learner who used very conspicuous communication strategies, who drew attention to the monitoring, and who did not fill his pauses, was judged to be a poor speaker and was, therefore, virtually ignored." (DeKeyser 1989: 115)

These views on the evaluation of performance involving reduction strategies could be summarised as is Table 2.5.

Table 2.5: The Evaluation of Reduction Strategies

	Functional or Message Reduction	**Formal Reduction**
Practised	Not observable; implicitly evaluated positively	*Either*: Potentially observable if result in interlanguage forms; possibly evaluated negatively, possibly neutral *or*: Not observable; implicitly evaluated positively
Not practised	Observed [Breakdown]; evaluated negatively	*Either*: Other strategies: e.g. approximation, restructuring; evaluated positively unless very hesitant *or*: Breakdown: evaluated negatively

A fundamental issue in the evaluation of compensation strategy use indicated by the table above is whether strategy use which can be identified is evaluated as part of an overall holistic impression of language proficiency as, in effect, a negative indicator of missing linguistic competence or recall (Ellis 1984), a kind of slumming which may lead to fossilisation at a low interlanguage level (the "Higgs view" Van Ek 1987: 56) or whether one evaluates strategy use positively as indicating flexibility, an ability to cope with the unpredictable (Haastrup 1986; RSA 1989; Goteborg undated). If one wishes to profile strategic competence as an aspect of proficiency, a position strengthened by the fact that "it is doubtful whether there are any forms of L2 (compensatory) strategic behaviour which do not have their analogue in communication involving only native speakers" (Kellerman et al 1987: 101), one is clearly interested in the latter approach.

The question then becomes, is there a criterion that one could use as the basis for assessment—or description of levels of proficiency?

Varadi (1980: 61) suggested the evaluation criterion: "Does the utterance convey the meaning which the learner actually wanted to communicate?" However, there is an obvious problem with this: how do we know what the learner wanted to communicate, and how do we know the learner

knows? This approach implies a model of communication as "an exchange of ideas mediated by language" (Brown 1990: 8) which has been called a "transmissions metaphor" (Rost 1990: 2–3). Attempts to operationalise this view of language often lead to tasks scored by information units (as for example in International Studies in Educational Achievement studies like Gorman et al 1988). Such approaches could be said to debase language to "the means by which information is shunted from one person to another…like sums of money or bags of oranges" (Smith 1985: 201) an idea "attractive to people who work with computers" (Brown 1990: 8) but one which fails to take account of the fact that "the transmission of information is never a terminal goal" (Sinclair 1985 personal communication). As Alderson puts the issue:

> "Surely another view of communication is at least imaginable, and plausible: namely that meaning is neither transferred nor extracted, but is created by participants, is negotiated in encounters and remains vague, undefined to the outside world but private for each participant and doubtless different for each. The "information gap" view of communication is at best partial, at worst wrong." (Alderson 1983: 89)

A way out of this cul-de-sac may be the broader definition of Strategic Competence as an aspect of proficiency proposed by Bachman. In relation to spoken interaction, such an ability which would include:

- the planning, execution and assessment of the achievement of communicative goals (Faerch and Kasper 1983);
- the ability to keep discourse on course through "challenging" for clarification (Burton 1980);
- the turn-taking and topic management strategies (Sinclair 1981, Kramsch 1986) which even advanced students often still have trouble with (Gotz 1977 cited in Faerch and Kasper 1983: 45);
- the cognitive strategies for framing ideas in discussion, formulating and evaluating hypotheses (Barnes and Todd 1977);
- the collaborative strategies for eliciting, commenting on and referring to other contributions (Barnes and Todd 1977), which have been grouped with the cognitive strategies as "cooperative strategies" (Wilkinson 1992);

plus:

- communication compensation strategies, both reduction strategies (Faerch and Kasper 1983) and propositional strategies (Kellerman et al 1987).

In this study the classification for Strategic Competence given in Table 2.6 was adopted. This matched a simplified version of the Faerch & Kasper/Bachman concepts of Planning, Execution, Assessment to a development from Tarone's (1980; 1981) distinction between Interaction and Production Strategies.

Table 2.6: Categories for Strategic Competence

	Reception	**Production**	**Interaction**
Planning	Framing	Rehearsing	
Execution	Inferring	Compensating	Turn-taking Cooperating Asking for help
Evaluation & Repair	Monitoring	Monitoring and self- correction	Asking for clarification Communication repair

Descriptors for these categories were (a) grouped from existing scales, though coverage there was meagre; (b) formulated from a reading of the literature, and (c) edited from transcripts of discussions of learner performances by teachers in the workshops described in Chapter 4.

Linguistic Competence. Linguistic Competence is another area which it is difficult to define, describe and scale to levels. The fundamental issue is to distinguish between *outcome specifications*, which would state necessary mastery (ACTFL) or typical mastery (Languages Lead Body) of certain forms on the one hand, and suggested *content specifications:* a less rigorous requirement in which certain forms are merely suggested as good things to teach at this level since, after all, "a syllabus is an idealised construct which serves as a reference for teaching" (Widdowson 1990: 127), and it only "defines a route which is to be negotiated, with a starting point, a destination and a journey in between. The specification of the itinerary, and indeed covering the journey itself, is subject to negotiation among the participants" some of whom, with definable specific (i.e. LSP) needs may prefer "package tours" (White 1983: 81), some of whom might profit from receiving or developing "maps" to wander with (Romiskowski 1981).

The latter approach is taken in sets of stand-alone language specifications (like *Waystage*, *Threshold*: Van Ek 1976, Van Ek and Trim 1990, 1991) and in content specifications attached to scales of proficiency (e.g. Eurocentres, RSA Modern Languages 1989) as well as in syllabuses of different kinds:

- any formal syllabus;
- a syllabus with different strands snaking or clustering around a formal core (Brumfit 1980; Page 1983: 296);
- one with cyclical levels repeating functions with exponents of increasing complexity (Breen 1987: 89);
- one which builds up to activities at the end of units (Paulston 1971; Rivers 1972, Littlewood 1981b, cited in Clark 1987: 104–7), an approach now common in many EFL coursebooks;
- the so-called "reversible" approach in which either the instruction or activity syllabus may lead (Allen 83, Dodson 1983, Brumfit 1984, Clark and Hamilton 84 all cited in Clark 1987: 104–7), since the activities call for a "contingent use of language" (Widdowson 1984: 123).

However, the current study is concerned, like ACTFL and the British National Language Standards, with the definition of *outcomes* rather than learning content or syllabus design. Since attempts to formulate grammatical competence at different levels of outcome specifications tend to fall into at least one of a number of different traps, some discussion of these is perhaps necessary. There appear to be four main issues, which are discussed in turn:

- a focus on mistakes, listing at each level mastery which is expected, i.e. errors which will not be tolerated;
- an employment of descriptors based on a *simple - to - complex* comparison;
- a focus on grammar and syntax, undervaluing the importance of learnt formulaic "chunks;"
- jargon-ridden and/or vague descriptor style.

Mistakes: ACTFL associates specific mistakes with different levels in the language-specific scales, information which tends not to be supported by Second Language Acquisition Research. Firstly, this very focus suggests that progress is a question of making fewer errors, whereas:

> "…the more the learner knows, the more likely he is to make errors. In other words: the simple utterance a learner succeeds in constructing on the basis of his early analyses of the input cannot possibly follow the rules of the target language for lack of the necessary (syntactic) devices. The beginning learner tends to rely on certain general principles that have little to do with any specific language, his first language, for example. Thus we are led to the paradox that, to a certain extent at least, the learner is more apt to make errors due to his first language knowledge the more he knows about the second language." (Klein 1986: 108)

The fact that inaccuracy increases in intermediate students is common knowledge and is very well documented in Fulcher's data (Fulcher 1993). The tension between attempts, crucial to language development, to increase the range and complexity of language handled in order to achieve the goals of more differentiated communicative tasks, and the retention of accuracy—given that the learner has limited processing capacity—has been highlighted by Foster and Skehan (1994).

Secondly, the mistakes highlighted at different levels tend to be just plain wrong. Developmental stages are about emergence not accuracy—about a "qualitative change in performance." The change may well leave gaps (which could be filled through teaching) as "a learner moves on without learning rules which he/she would in principle be able to process" (Pienemann 1992: 23–24). The late 1970s US "morpheme studies" (Dulay and Burt 1974; Bailey et al 1974; Hakuta 1974/8; Larsen-Freeman 1978, see Larsen-Freeman and Long 1991 for a review) claimed to provide evidence of a necessary order but, because of the fact that they provide only accuracy orders aggregated across people taking a test, very many SLA researchers consider the results as "of little theoretical or practical interest" (Ellis 1989a: 306–7). Instead, the Multidimensional Model of SLA developed during the ZISA project (Zweitsprachenerwerb italienischer und spanischer Arbeiter) (Meisel, Clahsen & Pienemann 1981; Clahsen 1985; Pienemann & Johnson 1987) has been used in relation to English-speaking learners by Ellis (1989a) and applied to profiling second language development in Australia (Pienemann and Johnson 1987; Pienemann and Janssen 1989; Pienemann 1992). The Multidimensional Model suggests that only some structures are developmental (generally those constrained by speech processing capability since they require departing from a "natural" or "universal" word order); others are subject to variation dependent on the orientation or style of the individual learner. The style shows itself in the balance he/she strikes in

allocating scarce processing resources between attention to form and attention to communication—a conclusion also reached from other directions (e.g. Hatch 1974: rule-learners / data-gatherers; Skehan 1986/1991a: pattern-making problem-solvers / chunk-accumulating memorisers).

In addition, Tarone (1983), Ellis (1985a, 1985b, 1987) and Skehan (Foster & Skehan 1994; Skehan 1995a, 1995b) have proposed that such learner performance styles are themselves variable according to context, and that accurate use of particular forms will be greatly affected by the conditions and constraints of the task being performed (see Variability above). Using accurate use of particular forms in an interview as a cut-off criterion between proficiency levels is therefore misguided.

Simplistic equation of "typical errors" with bands on a scale of proficiency on the basis of teacher instinct or committee consensus is dangerous. It distorts the measure and it has a negative wash back effect in encouraging teachers to count mistakes rather than monitor the quality of performance.

Simple - to - Complex: To avoid this pitfall, many scales employ descriptors which are based on a *simple - to - complex* comparison, despite the fact that they do not have much theoretical basis for such comparisons. The current portrayal of such *simple - to - complex* comparisons relates to concepts of Universal Grammar in general and the concept of *markedness* in particular:

> "The developmental sequences seem to reflect the internal complexity of the structure or the structural sequence to be learned, hence the degree of markedness. It seems that the unmarked or the less marked items are learnt early, the more marked ones later." (Wode 1984)

Thus markedness may define both the "directionality of difficulty" and the "degree of difficulty" (Eckmann 1977: 320): "areas of the target language that are more marked than the native language will be more difficult, depending on the relative degree of marking" (Spolsky 1989: 122 citing Eckmann).

Markedness originates from the Prague school in relation to phonology, and was carried over to inflectional morphology by Jakobson (McLaughlin 1987: 97). It can be related to Clark and Clark's (1978) complexity principle: that:

> "…complexity of thought tends to be reflected in complexity of expression, where "more complex" is reflected in the addition of morphemes, the addition of features or the addition of rules." (Rutherford 1982: 86)

Kellerman, however, sees markedness as psychological rather than linguistic: a structure is marked if there is a simpler way of saying the same thing, or if there are other meanings of the lexical unit which a native speaker would regard as more central (Kellerman 1979: 38 cited in Rutherford 1982: 90). Behind both interpretations of the concept is the idea that what is marked is further away from the "core grammar" of universal grammar as a whole and the "core grammar" of that language in particular.

This is the kind of thing scale writers have in mind when they are grappling with how to phrase things, whether they see the explanation as linguistic (like Wode) or psychological (like Kellerman). There are however at least two central problems. Firstly, there is a circularity of argument as Rutherford admits: is a construction linguistically marked because it is psychologically complex, or is it psychologically complex because it is linguistically marked? Secondly, the "developmental hypothesis" of markedness outlined above is only one of two hypotheses of the influence of markedness on SLA. The other is the "transfer hypothesis" (White 1987)—since Eckmann's original concept was an attempt to update the contrastive analysis hypothesis. The most obvious case of potential for transfer is where the native language has an unmarked form, and the second language a marked one (Ellis 1985: 206) but, as White demonstrates using data from Canadian immersion, there is no reason why marked forms should not also be transferred (e.g. "Un chalet qu'on va aller à" cited by White 1987: 277). White concludes:

> "…it appears that a developmental view (of markedness) cannot be maintained in its pure form, where the assumption is that the learner can disregard his or her previous knowledge and start from scratch." (1987: 278)

Thus, although the concept is useful in giving some underpinning to intuitions about complexity, it is of little practical use in formulating definitions of levels of proficiency because the concept is interpreted in many different ways and because while some structures appear to be *developmental* (i.e. acquired in a set developmental sequence others are *variational*.

Formulae and Scripts: Descriptions of linguistic competence tend also to underestimate the importance of linguistic knowledge stored lexically—what has been variously described as "routines and patterns, prefabricated chunks" (Krashen 1981: 99) "prefabricated routines or unopened packages." Widdowson cites a list of native speaker examples (1990: 91) which bear a striking resemblance to the kind of functional exponents found in notional/functional syllabuses, commenting that native speakers have hundreds of thousands of such lexicalised sentence-stems (1989: 132–3). The idea that such units are picked apart in analysis as a learner progresses (Krashen 1981: 99; Klein 1986: 77; Widdowson 1990: 96) may not necessarily apply to all students. This point, plus the fact that mother tongue speakers use such scripts and clichés all the time suggests that they are an aspect of foreign language competence at all levels—what people probably mean when they talk about idioms. Skehan suggests that although learners may display a concern for form and syntax when precision and creativity matter, language which has been analysed and assimilated may well be stored *lexically* not only as words but also as such chunks of speech, and that "when accessibility and time pressure are paramount a lexical mode of communication will be relied upon, which draws upon a capacious, well-organised, and a very rapid memory system" Skehan (1995b: 545). He adds: "but that given that such a system, not inherently focusing on rules, may hit problems, it is possible to "shift down" to a more rule-governed mode of processing…as the need arises" (Skehan 1995a: 8).

Knowledge of appropriate scripts and schemata relates to socio-cultural competence and when it is covered in scales tends to be divided between labels like "Appropriacy" on the one hand and "Range" on the other. There is also, however, a direct implication for descriptions of linguistic (meaning grammatical) knowledge, particularly with regard to accuracy. It is very probable that sticking close to "islands of reliability" (Dechert 1983: 184) and using the "canned speech" of a learnt repertoire provides a performance that gives a much higher impression of accuracy than a performance in which the learner is using the language creatively to express their thoughts. Such chunks of speech or "instances" may have been learnt as chunks in particular contexts and remained (or continued to be used) unanalysed; other may have been consciously constructed initially, but then automated through frequent use into a chunk for future processing (Schmidt 1992, cited in Skehan 1995b: 8).

Scales tend to overlook this with the result that the limited knowledge and operational coverage of a, say, *Waystage* performer, is equated with an erratic, unskilful, inaccurate performance (Trim 1978: 6) when this may well not be the case. When an elementary learner is operating in a familiar "discourse domain" (Douglas and Selinker 1985) for which they know the scripts they may achieve a reasonable level of accuracy. Once again, becoming "intermediate" or "advanced" can lead to more, not less, error.

Descriptor Style: The English Speaking Framework is one of the few scales of proficiency which offers a sub-scale for Linguistic Skills. The descriptors are built up by systematically, including a statement at each level taken from sub-scales for degrees of skill in features of linguistic performance. However, progression is achieved almost exclusively through word-processing qualifiers which makes the scale very difficult to read let alone use. The descriptor for Level 5, the middle band of the scale, for example, reads as follows:

> "Applies linguistic skills to moderate level tasks with adequate confidence and competence. Presentation of basic message is adequately adjusted to audience's knowledge of the language. Fairly frequent language lapses necessitate repair to capture detail and subtlety. Basic organisation of text is adequate, with a moderate range of cohesive devices. Uses a moderate range of styles but lapses of appropriacy are fairly frequent. Has a moderate range of language structures and vocabulary. Applies a moderate grasp of accuracy to communication and examination tasks. " (Carroll and West 1989: 58)

The word "moderate" is used as a key concept at this level and appears five times. The word "adequate," with a rather similar meaning, appears three times. Both words appear in two sentences (1st & 4th). The result is that if one strips away the jargon, the definition, though apparently taking account of seven aspects of proficiency, does not actually say much more than "He's okay,—just." In the level below (Level 4), "moderate" in the last sentence is replaced by "limited;" in the level above (Level 6) it is replaced by "good;" in the level above that by "very good." The result is that "adjacent definitions can appear at first sight to be almost identical to the non-specialist reader, and the whole scale can make your head spin" (North 1992a: 167–8). If you strip away the jargon, all that is really being said is that a Level 4 person is pretty limited, a Level 5 person is "okay" and a Level 6

person is "good." Since this is a 9 point scale one is left with the feeling that, although an attempt has been made to identify 7 aspects of communicative proficiency, which the testing community would refer to as *criteria* (c.f. McNamara 1995: 164), the formulations one is dealing with here and hence the decisions taken with reference to them are not criterion-referenced at all. The wording is norm-referenced around the middle of the scale in the same way at it is in examination rating scales.

There are two issues here, which are both taken further in Chapter 3. Firstly, as Skehan (1984: 217) has pointed out, this is not criterion-referencing. Secondly, adjacent descriptors which are distinguished solely by word-processing a different qualifier do not meet the need for "incisiveness" (Champney 1941) or "definiteness" (Thorndike 1904/12 in Engelhard 1991a) which is a requirement for a valid measurement scale.

As regards the second issue, the ESU is not alone. Even when one takes concrete steps to avoid the implicit norm-referencing criticised above, e.g. by isolating key features of performance at each level (Alderson 1991: 81), it can be very difficult to avoid slipping into making distinctions through qualifiers, particularly when trying to "spread" an aspect across a given number of levels. Alderson mentions this problem in relation to the formulation of descriptors for grammatical accuracy and pronunciation for IELTS, which had to have nine bands like ELTS, adding the related problem that in deciding which qualifier to use at which level one is faced with seemingly unanswerable questions like: "is *some* more than *a few*, but fewer than *several*, or *considerable* or *many*. How many is *many*?" (Alderson 1991a: 82).

In this study, there has been an attempt to select, edit or write short, transparent, descriptors which (a) avoid listing of specific features like ACTFL and the British National Language Standards, which (b) focus on one thing and try to describe a "definite" type of behaviour and which (c) as far as possible avoid depending for their overall meaning on purely on distinctions between qualifiers. An example for Grammatical Accuracy is:

> *Can use reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations.*

This particular descriptor was edited from statements at Level 2 of the RSA Modern Languages Scale, Band 4 on Carroll's 1980 interview scale, Level 4 on the Eurocentres Scale oral assessment grid, Level 1 on Shohamy's 1981 Hebrew Rating Scale and Band 4 on IELTS. It still uses a

qualifier *reasonably,* but this is used in an attempt to qualify the degree of accuracy referred to *in this descriptor;* it is not contrasted with *very* accurately at a level above, or *fairly* accurately at a level below.

Descriptors were edited or produced for Range (subdivided: Morpho-syntactic Range; Vocabulary Range); Accuracy (subdivided: Grammatical Accuracy and Vocabulary Control) and Pronunciation. The Range/Accuracy distinction is a common one in scales of proficiency which helps focus raters' attention on the complexity of the language used rather than just registering mistakes. It mirrors the Knowledge/Control distinction of Bialystok and Sharwood Smith (1985).

Pragmatic Competence. In her discussion of the Canadian DBP results, Schachter (1990: 40–41) argues on theoretical grounds for a fundamental distinction between Linguistic Competence and Pragmatic Competence, which Chomsky defines as "knowledge of conditions and manner of appropriate use in conformity with various purposes" (Chomsky 1980: 224 cited in Widdowson 1989: 130). Another definition of Pragmatic Competence is "the ability to use language effectively in order to achieve a specific purpose and to understand language in context" (Thomas 1983: 92). This interpretation of Pragmatic Competence fits with McNamara's view that it represents the beginnings of a model of performance. It is supported by Levinson's statement:

> "…to invoke Chomsky's distinction between competence and performance, pragmatics is concerned solely with performance principles of language use." (Levinson 1983: 3)

It is also related to Skehan's treatment of "ability for use" and, as will be argued below, to the possible rating criterion Fluency, which some teachers interpret so broadly as to include other, discourse aspects of Pragmatic Competence such as coherence, rhetoric structuring and topic development as well as propositional precision.

Propositional precision, being able to say what you want to, is related to the distinction between "sentence meaning" *Linguistic Competence* and "speaker meaning" *Pragmatic Competence* (Thomas 1983: 92 citing Leech 1983; Levinson 1983: 17 citing Grice 1957). As well as core (dictionary) meanings, words acquire meaning through negotiation in use. Accepting Searle's assertion that "one's meaning something when one utters a sentence is more than just randomly related to what the sentence means in the

language one is speaking" (Searle 1969: 45), there remains the problem of how much of this meaning is a core/prototype meaning—i.e. dictionary meaning and how much of it is pragmatic meaning—i.e. speaker meaning.

Neither Bachman and Palmer's 1982 study nor the DBP project shed much light on whether lexical knowledge should be considered part of Pragmatic Competence or as a part of Linguistic Competence (Harley et al 1990: 20) as in Canale and Swain, Van Ek and Bachman's later models. Schachter reports that Chomsky appears to include conceptual (i.e. *lexical*) knowledge in Pragmatic Competence. This was also the view taken at the time by Bachman and Palmer (1982) based upon intuitive interpretation of experience, that some learners just concern themselves with getting their meaning across through a combination of discourse and lexical skill, the kind of learners who have since been called *chunk accumulating memorizers* as opposed to *pattern-making problem-solvers* (Skehan 1986; 1989a: 36–7) as discussed in relation to variability. Schmidt's longitudinal study of Wes, a 33 year old Japanese artist on a 3 year stay in the US, for example, is a classic, extreme profile of a very successful, rhetorically expressive communicator, who hated classroom learning and proceeded by accumulating and memorising data—socio-cultural schema and conversational scripts and prefabricated chunks for different situations. These he used in conjunction with a very high level of (spoken) discourse and strategic competence to sustain highly effective, if idiosyncratic, performances which both he and the people he came into contact with considered perfectly adequate. He showed little or no sign of analysing linguistically the routines he used (Schmidt 1983: 150).

On the other hand, one could say that Wes's proficiency had little to do with lexis in the sense of dictionary meanings, vocabulary resources. He possessed a range of praxeogrammes (c.f. Ventola's (1983) flow charts of what might happen in a service encounter) together with routinised scripts to go with them and good interaction strategies. In terms of the pragmatic features Levinson (1983) focuses on, Wes seems to have had good discourse functions, good implicature (what one can reasonably take as a "given" in the context: Grice 1975), little textual cohesion but adequate coherence for spoken language. In terms of Widdowson's (1989) distinction between Analysability and Accessibility, Wes staked all on accessibility; he was no analyser, he was interested in Use not in Usage (Widdowson 1979). In terms of Brumfit's (1984) Accuracy/Fluency distinction which reflects the way many teachers think, having what has been called "intuitive reality"

(Fulcher 1993: 122) Wes was fluent, but hardly accurate. In terms of Skehan's (1995b: 552–3) distinction *undesirable* fluency (excessive proceduralisation due perhaps to using strategic competence to solve communicative problems) and *effective* fluency, Wes appears to be a pretty good example of the former, though it is worth recalling that that is apparently not how he or his conversational partners are reported as viewing the matter.

Fluency is a very popular category but also a problematic one because people think they know what it is, but it is seldom defined. As Elizabeth Ingram once put it:

> "There has been a certain amount of not very fruitful discussion in recent years about "fluency" from both the teaching and testing points of view. This is not surprising, for "fluency" is not a defined term, it is just a label, giving a very misleading impression of simplicity to that complex thing, command of spoken language." (Ingram E. 1968: 78)

If one takes the broad definition of Pragmatic Competence offered by Chomsky and Thomas (1983), then certain aspects of what is often understood under "fluency" would be considered Pragmatic Competence: the ability to *use* the language to express yourself and achieve what you want to (speaker meaning: Leech 1983). Other aspects, however, are clearly psycholinguistic. A higher level of competence leads to faster more efficient automatic access to knowledge as this is reorganised into integrated sub-routines (McLaughlin, Rossman and McLeod 1983: 136–7). It leads to procedural as opposed to declarative knowledge (Faerch and Kasper 1983) as smaller productions are unitised into larger productions (Kennedy 1988). Progress in procedural knowledge is progress away from the necessity for conscious controlled use towards automaticity (Bialystok and Sharwood Smith 1985).

Fulcher has reviewed the treatment of Fluency in existing proficiency scales, criticising the vagueness of the FSI definitions and the reliance of the Cambridge First Certificate (FCE) and Proficiency (CPE) on the concept of hesitation. He argues that unless speech is pre-planned, hesitations and reformulations will abound for native or non-native speakers, since it is in practice linked to forward planning and lexical choice (Fulcher 1993: 123–5).

In a study of Fluency ratings, Lennon confirmed Ingram's feeling that Fluency is a very global concept, concluding that Fluency represents "the speaker's ability to focus on the listener's attention on his or her message by

presenting a finished product rather than inviting the listener to focus on the workings of the production mechanisms" (1990: 392). He distinguishes two levels of looking at Fluency: (a) this kind of broad holistic impression used in statements like "He's fluent" or "Fluent User", which perhaps corresponds to the non-specialist focus on task completion (McNamara 1990, cited in Elder 1993) rather than as an aspect of competence, and (b) a narrow, more technical view of aspects of the performance. Lennon and Davies (1989: 167) both cite Fillmore's classic definition of "fluency" among native speakers, which could be glossed as follows:

- the ability to talk at length: *Narrower: psycholinguistic—related to "Flow." "Size" is sometimes referred to as an aspect of discourse competence;*
- the ability to talk in coherent, reasoned, and semantically dense sentences: *Middling: basically Pragmatic/Discourse Competence, including Coherence; and Propositional Precision;*
- the ability to have appropriate things to say in a wide range of contexts: *Broader: Pragmatic (use) but clearly also socio-cultural (scripts, schemata);*
- the ability to be creative and imaginative in language: *Broader: Pragmatic including Flexibility.*

Experience in the workshops described in Chapter 4 suggested that teachers had a range of interpretations of the term similar to Fillmore's above. Some, apparently influenced by Brumfit's (1984) basic Accuracy/Fluency distinction, adopted a very broad definition of Fluency which included everything to do with the communicative use of language, including strategies. Study of recordings of lengthy discussions of "Fluency" by two groups of native-speaker EFL teachers suggested that some teachers held several definitions of differing breadth simultaneously, as indicated in the following extract:

A "She was more fluent in the sense of—not fluent in the sense of being able to reach her aims better or communicate better. She was more fluent in the old sense of speaking more quickly. There weren't so many hesitations."

B "Oh yes, in that sense of Fluency, she was extremely fluent, and in that sense—"the ability to talk readily"—he was less so. She had tremendous *speed*." (Zürich, Adult Education, 21.3.94)

There seemed to be three main interpretations amongst teachers:

1. *Accessibility*: the ability to access knowledge and articulate a sustained flow in a stretch of speech. (Implies Size as well).
2. *Pragmatic or Discourse Fluency*. The above plus the ability to adapt the means of expression to the situation (*Flexibility*) finding ways of saying exactly what is meant (*Precision*) in well structured (*Coherence*) logically developed (*Thematic Development*) discourse.
3. *Communicative Fluency*. The above plus Interaction Strategies like knowing when and how to take the turn (*Turntaking*), the capacity to work with other people, commenting on the contributions of others (*Cooperating*: interpersonal), summarising points reached, framing issues to be addressed (*Cooperating*: ideational); checking understanding (*Asking for Clarification*) exploiting the interlocutor as a resource (*Asking for Help*), and, when necessary correcting misunderstandings (*Repairing*).

In the categories finally adopted in this study, the definition of *Fluency* is the first, base definition of accessibility or speech flow. The second interpretation comprises the rest of what has been included under Pragmatic Competence—still concerning the speaker's meaning only. The aspects included under No 3. above are considered Interaction Strategies. Turntaking, which could very well be considered to be the primary aspect of Discourse Competence applicable to speaking is therefore in this interpretation separated from other discourse aspects (size, coherence, thematic development). This is because it is identified with Burton's (1980) concept "discourse challenge:" taking the initiative in the sense used in Sinclair and Coulthard's (1975) discourse analysis model (Initiation, Response Feedback). It is considered to be an Interaction Strategy.

The decision to put Fluency under Pragmatic Competence cuts across the traditional competence / performance dichotomy used by linguists since Fluency is clearly related to performance. However, as was discussed above there are several arguments against continuing that tradition. Associating Pragmatic Competence with Use and Linguistic Competence with Knowledge or Resources allows one to keep Fluency (in the narrower sense of flow) together with the other elements in teachers' wider interpretation of the term (No 2 in the list above), which seems sensible in an approach designed for practitioners. The resulting set of categories for Linguistic and Pragmatic is presented in Table 2.7.

The following Fluency descriptor came out calibrated at what was subsequently interpreted as *Threshold Level*:

> *Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.*

Table 2.7: Categories for Linguistic & Pragmatic Competence

Linguistic	**Pragmatic**
Language Resources	*Language Use*
Usage	*Use*
What can be said (sentence/dictionary meaning)	*What people say (speaker meaning)*
Knowing a language	*Knowing how to use a language*

Scale Categories		
Range:	General, incl. scripts, formulae	Fluency
(Knowledge)	Morpho-syntactic	Flexibility
	Vocabulary	Precision
Accuracy:	Grammatical	Coherence
(Control)	Vocabulary	Thematic Development
	Pronunciation	

That descriptor was edited from a whole range of scales: English Speaking Union Band 4; Carroll interview scale 1980: Band 4; Eurocentres RADIO oral assessment grid Level 4; Wilkins 1978 (in Trim 1978) Band 3; ASLPR Band 1; Elviri et al (in Van Ek 1987) Band 3; RSA Modern Languages Degrees of Skill Band 1; Gothenburg University Band 1; Fulcher 1993 Band 1; ACTFL Intermediate-Mid; Ontario ESL Placement Band 4; Dutch Secondary Band 2. What is impressive is that in the preliminary subjective establishment of equivalencies between the source scales, based upon comparing wordings and video samples (when available) *all* of those levels on those scales had been identified as *Threshold*. except RSA Level 1. The only element from the RSA definition reflected is *Will be extremely hesitant, pauses and reformulations to be expected*. (Royal Society of Arts 1989: 6) which could also apply to a lower level.

For nearly all the descriptors which survived the weeding out, editing and calibrating process there is a very high degree of agreement between the initial, provisional assignment to a level and the final calibrations, but no

other category has a relationship quite so high as for the Fluency sub-scale. For Fluency the RSA reference is the only one out of 29 source references which is "misplaced"—that is to say not calibrated at the level it was provisionally placed in the descriptor pool. This degree of consistency in the interpretation of the descriptors for Fluency would seem to reinforce the findings of Lennon (1990: 392) that in terms of rater perceptions, fluency is a construct very central to language proficiency, and those of Fulcher that a Fluency scale can operate with a relatively high degree of reliability across raters and tasks (Fulcher 1993: 293).

Socio-cultural Competence. Socio-cultural competence is generally recognised as being extremely important in curriculum development and is given prominence, for example, in the Canadian Communicative / Experiential Syllabus, which cites Saville-Troike's (1983: 131–2) statement:

> "The concept of communicative competence must…be embedded in the notion of cultural competence; interpreting the meaning of linguistic behaviour means knowing the cultural meaning of the context in which it occurs" (c.f. Lessard-Clouston 1992: 328).

Attempts to write scales for socio-cultural competence have, however, been notoriously unsuccessful, as both the writers of the culture scale attached to the 1982 provisional version of the ACTFL Guidelines and the attempt by the American Association of Teachers of French (National Commission on Proficiency; Committee on Cultural Competence) admit. In a critique of the latter Byram and Zarate (1994: 20–24) comment that the American Association of Teachers of French (AATF) have proposed a 4 band scale, rather than the 11 band ACTFL scale used for language assessment, but that "the decision about what constitutes each level in terms of abilities and knowledge remains unexplained" (ibid: 20). The proposed scale has 3 sub-categories: socio-linguistic ability, knowledge of cultural areas, knowledge of cultural analysis, with each level defined in terms of what the learner "can demonstrate". Byram and Zarate (ibid: 24) conclude that the descriptors themselves demonstrate the difficulty of writing socio-cultural objectives, being clearly open to several interpretations. Kramsch's verdict on the AATF scheme is:

> "These first steps towards a classification of cultural performance and competence show how risky any attempt at developing a national instru-

> ment for evaluating cultural competence is bound to be." Kramsch (1989: 6)

Socio-linguistic competence (often called Appropriacy), one of the three AATF categories, is a relatively common category on scales of proficiency. It concerns knowledge of rules of style, directness, and appropriateness. Socio-pragmatic failure is caused by different beliefs about rights/imposition (e.g. physical closeness, power and turn-taking conventions, mentionables/taboos as opposed to linguistic mistakes or pragma-linguistic failure—incorrectly/inappropriately mapping form to function in speech acts (Thomas 1983). The use of wrong "behavioremes" (Oksaar 1992) appears to be judged far more severely than L2 errors and to be a far greater barrier to international understanding. Socio-linguistic competence in this sense is thus concerned with the choice of language which is appropriate to the relationship between the participants; the focus could be said to be on the people.

Another aspect of Socio-cultural Competence concerns the question of what patterns of moves occur, and so may be deemed appropriate, in particular texts and situations; here the focus could be said to be on the setting rather than the people. Knowing what usually happens and what *might* happen in different encounters (having praxeogrammes) could be regarded as another aspect of Pragmatic Competence except for the fact that all such schemata are bound by socio-cultural conventions. With regard to Service Encounters (one of the categories used for Communicative Activities) "the participants create the social process by following the different paths in the tactic pattern of the schematic structure of service encounters" (Ventola 1983: 247). Knowing what kinds of text patterns occur in speech or writing involves Socio-cultural Competence, since such encounters and text patterns take different forms in different genres in different societies. The importance of such schemata both in terms of expected content (*content schemata*) and the rhetoric organisation of that content (*formal schemata*) is also recognised in reading research (Carrell 1987; Carrell & Eisterhold 1983). A praxeogramme for a situation may also be called a *Script* or *Scenario* which is defined as a "mentally stored representation of a familiar event like visiting a restaurant" (Murphy & Cleveland 1991: 150), which Davies sees as socio-cultural, and of which as he says there are always more to be collected, by native and non-native speakers alike (Davies 1989: 168–9). Such knowledge is usually referred to in existing scales of language proficiency as knowledge of socio-cultural conventions.

Finally, inter-cultural skills are an aspect of Socio-cultural Competence not found in any of the scales analysed. The curriculum aim of "intercultural skills" is to create "150% persons" who understand (empathy) find value in and have positive sentiments towards (favourableness) both cultures. They are effective at interaction in both cultures and at acting as "mediating persons" (Saltzman 1986 in Lambert 1993b: 191). Byram and Zarate propose such an "intercultural speaker" rather than the native speaker as the norm for comparison in developing descriptors. The concept of intercultural skills is not new. Loveday (1982: 50–51) cites a list of skills for intercultural communication drawn up by Seeyle 1977:

- the sense of culturally conditioned behaviour (handshake)
- the interaction of language and social variables
- conventional behaviour in common situations
- the evaluation of statements about a society
- cultural connotations of words and phrases
- skills to locate and organise L2 cultural info from e.g. library
- demonstration of empathy and curiosity towards L2 culture.

If first steps towards a classification of Socio-cultural Competence are "risky" (Kramsch 1989: 6), first attempts at scaling are doubly so. None of the scales used as sources in this study treat Socio-cultural Competence well. Attempts to produce progression tend to produce results like the following example for adjacent levels of the Gothenburg scale, in which the progression is created just by adding or changing qualifiers. As will be discussed in Chapter 3 in relation to the essentials of a valid measurement scale, such differences do not in themselves provide much "incisiveness" (Champney 1941) or "definiteness" (Thorndike 1904/12 cited in Engelhard 1991a).

Shows good awareness of cultural differences.
Shows some awareness of cultural differences, level of directness usually correct
Shows some elementary awareness of cultural differences. Sometimes too direct

Apart from these problems of formulation, there is also the question of whether socio-cultural competence *can* be put on the same scale as language competence. Bachman and Palmer (1982) found evidence for the fact that pragmatic/grammatical competence formed one trait, and that what they called socio-linguistic competence, defined as "distinguishing of registers,

nativeness, and control of non-literal, figurative language and relevant cultural allusions" formed another (Bachman and Palmer 1982: 450). From their definition it would appear that they are using the terms "socio-linguistic" in a broad sense; their definitions would presumably include knowledge of socio-cultural conventions.

Pollitt and Hutchinson (1987) also had problems with socio-linguistic competence, operationalised as the criterion called Appropriacy, in their Rasch model scalar analysis of different aspects of competence on different writing tasks. Socio-linguistic "items" misfitted in the Rasch analysis, they just did not belong in the same scale as "items" on discourse and linguistic competence.

In this study, the preliminary workshops with teachers on socio-cultural items (undertaken by a colleague in the team) also proved to be the most problematic. This may have been partly due to the nature of the informants, who were university *Romanistik* students (i.e. students of French language and culture) and Gymnasium teachers of French, rather than English teachers as with the other workshops. However, the informants' inability to separate descriptors supposed to be about Knowledge of Socio-cultural Conventions from descriptors on Socio-linguistic Appropriacy may reflect Bachman and Palmer's use of Socio-linguistic to include more than just appropriateness, directness, style and register. Their limited ability to organise the descriptors into vertical categories by level (Low, Middle, High) may reflect the problems referred to in relation to the Gothenburg definitions above.

Despite these problems, what appeared to be the twelve clearest descriptors on Socio-cultural Competence were included in the survey. Misgivings were confirmed when all but two items misfitted wildly in early analyses and so the category was dropped. The two descriptors which fitted were the following:

Can adopt a level of formality appropriate to the circumstances
Can use simple everyday polite forms of greeting and address

Each was saying something noticeably "definite", something "incisive" so they were reclassified to Flexibility and to Conversation (one of the categories on the Activity side) respectively, kept in the analysis and calibrated.

Accommodation. The need for accommodation on the part of the interlocutor is another problematic category. Accommodation is a reaction of

the interlocutor to the learner's competence rather than strictly speaking, an aspect of that competence, but it is an indicator of proficiency related to Language Use. Labelling the phenomenon positively from the point of view of the learner rather than interlocutor, the term "Independence" has been adopted from Carroll (Carroll B.J. 1980).

There is little doubt that native speakers do on some occasions adjust to take account of learner competence, and that the need for this adjustment decreases as the learner gains proficiency. Varonis and Gass (1982: 117) emphasis that (in giving directions asked for by real or fake NNSs) native speakers tend first to confirm they have understood what the non-native speaker wants. This accords with Long's view that native speaker modifications are primarily *interactional* rather than linguistic and that they "serve to provide input that is well-formed, a sort of linguistic and conversational cocoon for the neophyte second language acquirer" (Long 1983a: 186) and that they are aimed at "two main ends: (1) to avoid conversational trouble, and (2) to repair the discourse when trouble occurs" (Long 1983b: 131). The hardest evidence that naive native speakers *do* make substantial *linguistic* adjustments and corrections appears so far confined to Japanese (Shorted 1993; Berwick and Ross 1993). Rost (1990: 163) and Ross (1992: 177) list features of native speaker accommodation which include grammatical simplification in the classroom context and in an assessment interview context respectively.

Ross (1992: 183) goes as far as to suggest that the inverse correlation between proficiency and extent of accommodation (a point also noted by Long 1983b: 186) could provide a "useful metric for assessment" at the lower end of the proficiency continuum where most learners are situated. To elaborate Ross's point, accommodation could (a) be recognised as a feature affecting holistic judgements—that it could in fact in reality be *the* criterion for many interviewers (1992: 181–2), (b) be used as a holistic assessment device at lower levels or (c) be a category in an analytic scale. Ross and Berwick (1992) show that an analysis of accommodation exponents predicted 9 out of 15 ILR oral proficiency results of 1+. Ross and Berwick also demonstrate the danger to the validity of interviews of teachers' lack of awareness of how and when they accommodate, often when it is unnecessary and counter productive (Ross and Berwick 1992: 170). This confirms the general findings in accommodation research: that people accommodate towards where they think the other person is, or what they think the other person needs and often "overshoot" (Giles et al 1991: 13) in

a patronising manner which sometimes produces tragi-comic results. Such psychological or subjective convergence (as opposed to real or objective convergence) can even lead to situations straight out of sit-coms: "where sojourners in foreign climes actively converge over time towards the (often ill-conceived) convergent attempts of individuals from the host community towards them!" (Cohen and Cooper 1986 cited in Giles et al 1991: 15).

These problems do not, however, alter the fact that the interlocutor is a very significant factor in performance since communication in casual conversation is achieved by negotiation between at least two parties (Savignon 1983: 8–9; Kramsch 1986: 367). Scales of language proficiency often try and take account of this fact with concepts like: "native speakers used to dealing with foreigners attempting to speak their language" (FSI Level 1) which reappears in "persons…used to speaking to non-native speakers" (ASLPR: 0+) "native speakers in regular contact with foreigners attempting to speak their language" (ASLPR 1) and "sympathetic interlocutors" (ACTFL Intermediate-Low). However, it seems preferable to treat the issue as one affecting all levels, rather than as a problem confined to elementary proficiency, as in the FSI "family" of scales. This is the approach taken originally by B.J Carroll (1980) with his criterion "Independence" which reappears laconically in the CASE scales as "Interlocutor support: not / occasionally / frequently / continually required" (Cambridge Assessment of Spoken English Mnemonic scale 1993: bands A–D); or more extensively defined in the RSA (1989) Modern Languages.

Three categories for Independence were used in analysing the source scales and carried over into the descriptor pool: Need for Interlocutor Adjustment, Need to Get Clarification (obviously related to each other) and Need for Support, the flip side of the interaction strategy Asking for Help. In the Rasch model analysis of the survey data, however, these categories all proved highly unsuccessful, as is discussed in Chapter 5.

Categories for Communicative Activity

> "Since the sixties it has become increasingly clear that a simple classification of proficiency as the "four skills" of listening, speaking, reading and writing is inadequate, particularly for curriculum development and testing." (Stern 1983: 347)

However, the framework of the four skills plus maybe *Use of Language* and *Culture* remains popular with schools, publishers and examining boards, is understood by students and avoids the problem of deciding is what to put

in its place. So, "holding onto nurse for fear of something worse" the vast majority of scales of proficiency stick to a four skills model with or without a global scale. This section explores the question of whether an alternative method of classification could be used. First a grouping by functional categories is considered in an attempt to see whether there are "macro" functional categories which can be related to theory. Then the areas of overlap between the four skills are discussed, with different groupings being proposed. This is followed by a discussion on the fluidity of the categories and a consideration of the advantages that such an approach might bring over one which sticks to the four skills.

Macro-Functions

A functional rather than skill orientation was pioneered in the 1970s by the three LSP projects mentioned previously: ELTDU; IBM France, and the Canadian Language Selection Standard. The problem is that any attempt to follow a functional, task, activity classification prompts the comment "why *this* set of functions, uses, rather than any other?" (Davies 1990: 41 discussing the ELTDU scale). This suggests that a communality of functional purpose or social goal needs to be established before a single scale is justified (Spolsky 1986: 154; 1989: 65). The 1983 version of the Eurocentres Scale of Language Proficiency, (Johnson and Scott 1984) was organised according to the macro-functions shown in Table 2.8. This organisation was selected after an analysis of the results of a survey of perceived needs with a 30% representative sample of Eurocentres UK students learning English.

There is a certain influence from Halliday in the "Communication" scales, the first two of which more or less split the *ideational* aspect of language into its two elements (a) experiential; (b) logical (Halliday 1978: 187), the third concentrating more on the *interpersonal*: "use of the language to express social and personal relations" (Halliday 1973: 41), "the expression of our own personalities and personal feelings on the one hand, and forms of interaction and social play with other participants in the communication situation on the other hand" (Halliday 1973: 66), with the *textual* aspect which "enables the speaker to organise what he is saying in such a way that it makes sense in the context and fulfils its function as a message" (Halliday 1973: 66) involved in all three.

The approach, whilst valuable for the elaboration of content specifications (tasks, functions, structures), did not prove practicable for use in a Eurocentres context in scalar form. The reason was that all three aspects

(ideational, interpersonal, textual) were interwoven in any situation, requiring the use of three scales simultaneously for any observation: "ideational content and personal interaction are woven together with and by means of the textual structure to form a coherent whole" (Halliday 1973: 109), and the scales themselves were constructor-oriented (saying what the learner could do) rather than assessor-orientated (Alderson 1991a: 72–4).

Table 2.8: Categories in the Eurocentres Scale

	Communication	**Understanding**
Spoken	Asking for and exchanging information (*Information*) Expressing opinions and making judgements (*Opinion*) Establishing and maintaining social relationships (*Social*)	Listening to authentic texts
Written	Asking for and exchanging information (*Information*) Expressing opinions and making judgements (*Opinion*) Establishing and maintaining social relationships (*Social*)	Reading authentic texts

Some other curriculum systems (not usually scales of proficiency) also use this type of classification, e.g. the Australian National Assessment Framework for Languages at Senior Secondary Level (Years 11 and 12 of Australian high schools) taken from the Australian Language Levels (ALL) (Tuffin 1990):

1. Establishing and maintaining relationships and discussing topics of interest
2. Participating in social interaction related to solving a problem, making arrangements, taking decisions with others, and participating in transactions to obtain goods, services and public information
3. Obtaining information (a) by searching for specific details in a spoken or written text (b) listening to or reading a spoken or written text as a whole and then processing and using the information obtained

4. Giving information in spoken (monologue) or written form on the basis of personal experience (talk, essay, instructions)
5. Listening, reading, viewing and then responding personally to a stimulus
6. Creating (a story, dramatic episode, poem, play)

This is a simplification of the approach of the ALL communicative syllabus (Clark 1987) shown in Table 2.9.

Table 2.9: Categories in the Australian Language Levels

Dimensions	**Modes**
-establish and maintain interpersonal relations	-conversation and correspondence
-acquire information from more or less public sources	-giving information in speech or writing in a public form -listening and reading for information & pleasure
-listen to, read, enjoy and respond to creative and imaginative uses of target language	-creating imaginative text (some learners only) -listening and reading for pleasure (responding to more imaginative uses of language)

Relationships between the Four Skills

Other applied linguists have also tried to get away from the four skills and emphasise ways in which they are integrated. Carroll (1983: 102) cites Spearitt (1979: 7) that: "there is no support for an expressive skill category incorporating both writing and speaking, but there appears to be some support for a receptive communication skill embracing reading and listening."

Certainly in literature on reading and listening (e.g. Matthews 1990b; Rost 1990, 1993; Brown 1990; Buck 1991; Brindley and Nunan 1992; Urquhart 1992; Moran and Williams 1993) one is struck by the similarity of the listening and reading processes (Rost 1990: 18), the similarity of the role of bottom up/top down processing (Rumelhart 1977), of the use of content and formal schemata (Carrell 1983; 1987; Carrell and Eisterhold 1983) in inferencing (Brown and Yule 1983: 251ff).

With speaking and writing the position is more complex. That there is an overlap is undeniable:

> "The idea that spoken language is formless, confined to short bursts, full of false starts, lacking in logical structure etc. is a myth—and a pernicious one at that, since it prevents us from recognising its critical role in learning. It arises because in writing people only ever analyse the finished product, which is a highly idealised version of the writing process; whereas in speech they analyse,—indeed get quite obsessed with—the bits that get crossed out, the insertions, pauses, the self-interruptions, and so on." (Halliday 1989: 89–90)

Halliday gives three reasons for the focus on the difference: (a) value systems of literate cultures (b) the fact people went wild when with tape recorders became available, and (c) the fact that the analysis process started with academic seminars, which are atypically disjointed, because those taking part are working out their argument as they go along.

Halliday concludes "in a literate society, the functions of language are shared out between speaking and writing; there is some overlap, but by and large they fill different roles" (1989: 100). What is done by each is "partly a ritual matter, a form of social convention" (1989: 93).

Precisely what is the role of written language in Western culture is disputed: there are those who consider that:

> "It seems reasonable to suggest that, whereas in daily life in a literate culture, we use speech largely for the establishment and maintenance of human relations (primarily interactional use), we use written language largely for the working out of and transference of information (primarily transactional use)." (Brown and Yule 1983: 13)

There are others who feel that:

> "Speech is better than writing for communication. The question is essentially of resources. Since language is basically fallible as a means of information transmission, it needs to draw on everything it can to achieve communication effectively…. And language that is spoken has more resources than language that is written…But writing is infinitely more efficacious than speech in another respect. It is superbly more potent in creating worlds.…The danger is that the information-transmission emphasis can lead to an almost exclusive perception of writing from the perspective of the reader, rather than from the writer's point of view" (Smith 1985: 205; 210)

One sees account taken of the humanistic educational point Smith is making on the concentration on imaginative uses of language, on what has been labelled *autonomous* competence as opposed to communicative competence (Canale 1984) in the ALL dimensions and modes listed above. Autonomous tasks can be seen as context-reduced, requiring a form of CALP: cognitive academic language proficiency: Cummins 1979; 1980; 1983). Canale goes on to suggest that in autonomous tasks there will be less emphasis on strategic and socio-linguistic competence, and more on discourse competence (i.e. knowledge of the formal schemata associated with the genre) and grammatical competence since the work will be read, in Halliday's terms, as a finished product.

Communicative Activity and Supra-genres

Breen and Candlin (1980: 92) posited three underlying abilities: "Interpretation," "Negotiation" and "Interpretation," which they considered to be not primarily linguistic, but which Brumfit (1984: 69–70; 1987: 26) developed into "Comprehension," "Conversation or Discussion" and "Extended Speaking" (with Extended Writing as a supplementary category appropriate on some occasions) which he described as "independent modes of behaviour" (1984: 70). Brumfit proposed integrating the three abilities proposed by Breen and Candlin with linguistic behaviour in this way in order to reclassify the four skills more in line with "common sense assessments of what we do with language" which integrate activity with communication and could be used as goals towards which to orientate classroom activities.

Alderson and Urquhart (1984: 227) consider the issue more from the point of view of the kinds of text university students are confronted with, and propose the categories shown in Table 2.10. The categories are a mixture of supra-genre (e.g. Correspondence) and language activity (e.g. Lecturing):

Table 2.10: Categories for University Students

	Dialogue	**Monologue**	
		PRODUCTIVE	RECEPTIVE
Spoken	Conversation	Lecturing	Listening
Written	Correspondence Telexes	Writing	Reading

Swales (1990) considers these issues from a *genre* angle. Genres, which he concedes (1990: 33) remain "a fuzzy concept" and which he defines as "classes of communicative events which typically possess features of stability, name recognition and so on" (1990: 9) are used in communication by socio-rhetorical networks which he calls "discourse communities" for the "communicative furtherance of (*a*) set of common goals." Discourse communities are characterised by the fact that established members are familiar with the particular genres involved to do this (Swales 1990: 9–10). Swales proposes (1990: 55) that genres can often be grouped into "supra-genres" or "convenient multigeneric generalisations" like *Letter*, and that different, relevant genres of letter can then be introduced in the syllabus since as units for syllabus organisation, genres are about the right size (1986) and since, because genres are socio-culturally defined entities he feels learners cannot be expected to acquire out of thin air genres just because they acquire linguistic knowledge. In Davies' (1989: 168–9) terms, language knowledge or proficiency by itself is not a lot of use without the socio-cultural knowledge of the scripts and schemata associated with the different genres in which one has to operate.

Furthermore, Swales argues that certain types of language use can be regarded as *pre-generic* and common to all societies: casual conversation or "chat", and "ordinary" narrative storytelling (Swales 1990: 58–61). The former "chat", is *interactive* with short turns, its coherence provided through the way participants weave their contributions together. It can be related to Cummins (1979; 1980) concept BICS (Basic Interpersonal Communications skills), tending to low cognitive complexity and high contextual support (implicature). It can be considered to underlie all the genres of more specialised communicative interaction developed in different cultures (Swales ibid). The latter, storytelling, is *productive,* often prepared, rehearsed (or redrafted), its coherence provided in the text by the speaker/writer. It can be considered to underlie literacy and can be related to Canale's (1984) concept "autonomous competence" (as opposed to communicative competence), tending to high cognitive complexity and low contextual support.

The distinction between *interactive* and *productive* language is also related to the distinction made by Brown et al (1984) between short and long turns. "Long turns which are used to transfer information—to recount an anecdote, justify a position, give instructions about how to take some medicine, describe a route—demand skill in construction and practice in execution." Very young children do not attempt long turns, and young people in general

(and some adults) have great difficulty organising them when under communicative stress (Brown et al 1984: 15). Storytelling, production, long turns, then create an inverse *receptive* role as an auditor/recipient.

North proposed regrouping skills under the three headings: Reception, Interaction and Production, which reflect the three Alderson and Urquhart categories Receptive, Dialogue, Productive (North 1992a), adding a fourth category Processing (North et al 1992) or Mediation (Council of Europe 1996). These distinctions relate to those made by Breen and Candlin, Brumfit and Swales as laid out in Table 2.11. Although each author uses a very different title to describe the area being discussed, there is a certain similarity in the distinctions being made.

This kind of distinction has also been adopted in the Canadian Communicative / Experiential Syllabus (Tremblay 1990: 20). This also employs a three category distinction between Comprehension, Negotiation, and Production. In this Canadian division, as in the suggestion in North 1992a and its development in North et al 1992, the production of all written texts is placed under *Production*.

Table 2.11: Categories as Alternatives to the Four Skills

Underlying Abilities	**Major Activities**	**Pre-genres**	**Macro-skill**	**Activities**
Interpretation	Comprehension	[Listening to story-telling]	Reception:	Viewing Listening Reading
Negotiation	Conversation	Casual conversat-ion	Interaction:	Conversation Transaction Discussion
Expression	Extended Speaking/ Writing	Story-telling	Production:	Presentations Letters Reports
			Processing:	Listening and note-taking Reading and synthesizing Interpreting Translating
Breen and Candlin 1980	Brumfit 1984	Swales 1990	North 92	North et al 92

Alternatively, the production of text within the context of a slowed down interactive dialogue could be placed under *Interaction*, as in Alderson & Urquhart. There are arguments for including interactive writing under *Production* or under *Interaction*. Examples of this kind of writing would be correspondence (written conversation) or filling out a form or questionnaire (written structured interview). In whichever place interactive writing is put, the kind of coherent longer spoken turns or embedded monologues represented by narrative and extended description, which in speaking are embedded in the shorter turns of a conversation / discussion (Jeffersen 1978: 220), and which appear very frequently in personal letters (Swales 1990: 61) would be put under *Production* as in the Canadian scheme. Very similar language is involved whether the story is told or the town is described in speech or writing.

The main argument for including interactive writing under *Interaction* is that a great deal of the language used in this way is identical or at least very similar in speech and writing. Most learners find reinforcement by the written word useful if not essential and it could be argued that it is artificial to separate learning speaking from learning writing in this way for the literate second language learners one finds in countries like Switzerland. In addition, most interactive writing situations are tolerant of some error and confusion and have some contextual support. There is usually an opportunity to use interaction strategies like asking for clarification, asking for help with formulation and for putting right misunderstandings with repair strategies. Finally, the requirement to produce coherent text developed in a logical manner according to the conventions of the genre involved is negligible in this type of writing. On the other hand, autonomous competence (Canale 1984) is an aspect important in both written and spoken *Production*.

The main argument in favour of putting interactive writing in *Production* is both pragmatic and theoretical. In profiling achievement, drawing a distinction between ability to write letters in a correspondence and circulars (to which no reply is expected) could be seen as superfluous: in both cases it is primarily production strategies that are required, since there is no opportunity to negotiate meaning or to get help at the actual time of expression. From a theoretical point of view, one can also argue that "if dialogue has primacy over monologue, it is but a small step to seeing monologue as a specialised form of dialogue between the writer or speaker and the reader of listener" (Hoey 1983: 27). Hoey here is arguing that if, as argued by Swales (1990: 58–61) "chat" (*Interaction*) and "ordinary narrative storytelling" (*Pro-*

duction) can be considered to be pre-generic then it is a small stop to seeing the interactive mode as primary and so underlying everything. Winter (1977) and Hoey (1983) use a technique paraphrasing monologue into dialogue in the form of implicit questions and answers to elicit the clause relations which bind monologue discourse together, sometimes without any salient linguistic linking devices such as conjunctions, subordinators and lexical signalling. They would argue that *all* text can be paraphrased in this manner, as in the example given in Table 2.12.

Table 2.12: Text as Dialogue

Original Monologue	**Dialogue structure eliciting clause relation**	
Mr Wilson won many middle class votes in the election. He appealed to scientists and technologists to support his party	D:	Mr Wilson won many middle class votes in the election.
	Q:	How did he achieve this?
	D:	He appealed to scientists and technologists to support his party

Source: Winter 1977 in Hoey 1983: 29

If virtually all monologue can be represented as implicit dialogue in this manner, then the primary distinction could be best seen as being purely one of length of turn. An article, a novel, an 18 page love letter: these are all long turns requiring at least a minimum of discourse structure (clause relations) which provide a "covert interaction whereby (*the writer*) anticipates the likely reactions of an imagined reader and negotiates with him as it were by proxy" (Widdowson 1984: 88). To single out the love letter alone as being interactive seems simplistic.

A distinction between *Interaction* and *Production* based upon length of turn rather than text-type (e.g. letter) would in practice leave form-filling (written structured interview), postcards and short formulaic notes and messages, where the language is in effect written speech, under *Interaction* and put extended spoken and written turns under *Production*.

As with Tarone's (1983/1981) division of performance strategies into interaction and production strategies, the argument is that although pure *Reception* may be involved in *Interaction*—for example when, due to temporary shifts in discourse role in a discussion the learner is spoken across and becomes the "overhearer" of a dialogue or becomes the "addressee" of a monologue rather than an interactive participant (Rost 1990: 5), and al-

though conversely pure *Production* may also be involved in *Interaction*, particularly in structured situations like formal meetings, this does not alter the fact the processes involved in *Production* are essentially different to those involved in *Interaction*. In *Production*, more coherent, more prepared discourse is generated. With some practised speakers the product is virtually indistinguishable from their written prose. Oral examinations increasingly take account of the distinction between *Interaction* and *Production*. Some examinations evaluate the different types of discourse on different test occasions (Shohamy, Reves & Bejarano 1986), some have an interview structured to get samples of both interaction and spoken production (e.g. International Certificate Conference (ICC) Stage 3) and even small group interaction oral tests can be structured to generate samples of both types of spoken language (North 1991, 1993b).

A further advantage of such an approach based on genre and supra-generic categories is that genres are socio-cultural units, the acquisition of genre skills entails the activation of appropriate *content schemata* and acquisition of the appropriate *formal schemata:* principles of discourse organisation (Carrell 1983, 1987; Carrell and Eisterhold 1983) appropriate to the genre (Swales 1990: 9–10). Service Encounters (Ventola 1983), for example, could be regarded as a "convenient multigeneric generalisation" as could formal (chaired) discussions, presentations, reports, or interviews.

Cross-referencing Macro-Function and Mode

Another way of looking at this issue is to cross reference macro-functions (e.g. Transactional) to modes (e.g. Interaction) to produce "convenient multigeneric generalisations," or communicative activities (e.g. Obtaining Goods and Services) as in the cells of Table 2.13. Each cell in Table 2.13 can be elaborated for spoken language and for written language with one exception, *Discussion*, since written discussion (though it might be argued that there are interactive aspects to it) would be absorbed into the Production category *Presenting a Case*. The sub-categories in Table 2.14 emerged through an interplay between the work of the Council of Europe authoring group, and experimentation with categories in the workshops with teachers described in Chapter 4.

In Table 2.14 the Production categories *Describing, Narrating and Interpreting Experience* and *Presenting a Case* appear for both spoken and written language. Such production often takes the form of spoken monologue embedded within the interaction.

Table 2.13: Macrofunction Cross-referenced to Mode

	Reception	**Interaction**	**Production**
Transactional Language Use	Understand-ing Information-carrying Text	Obtaining Information and Services	Presenting Information
Creative, Interpersonal Language Use	Understanding Fictional Text	Maintaining Social Rel-ationships	Describing, Narrating and Interpreting Experience
Evaluative, Problem-solving Language Use	*(Merged with info-carrying texts)*	Discussion	Presenting a Case

Not all subcategories in Table 2.14 are described at all levels. *Service Encounters* and *Form-Filling* for example, are considered primarily a concern for learners up to about *Threshold Level*, whereas *Negotiating* is probably a skill first feasible from *Threshold Level*. Both these hypotheses were borne out in the calibration of descriptors.

Table 2.14: Categories for Communicative Activities

	Interaction		**Production**	
	Spoken	**Written**	**Spoken**	**Written**
Transac-tional	Service en-counters Information exchanges Interviews Telephone transactions	Form-filling Notes & messages Formal letters	Formal pres-entations	Formal reports
Creative, Inter-personal	Conversation	Personal letters	Describing, narrating and interpreting experience	Describing, narrating and interpreting experience
Evalua-tive, Problem -solving	Discussion Negotiating Formal meet-ings	- -	Putting a case	Putting a case

As discussed above, written language production is often associated primarily with transactional use, the transmission of content (e.g. Brown and Yule 1983: 13) but it can in fact be argued that written language is inferior to spoken language for the transmission of information. The business world appreciates the value of a spoken rather than written *Presentation*. On the other hand Production can harness the world of the imagination and present it either in writing (a magazine, an anthology) or in speech (e.g. a play, a video), so helping to create literacy (Smith 1985). The argument is made mainly in relation to mother tongue, but there are implications for language learning.

Purity of Categories

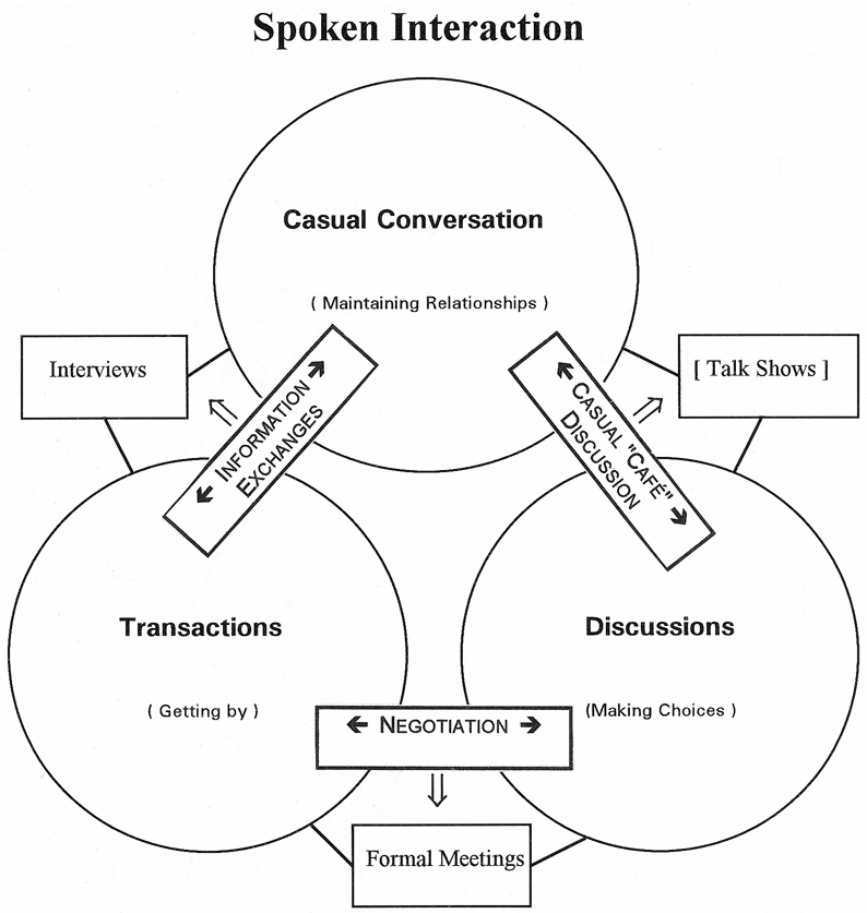
As with all sets of categories, which are but "conceptual artefacts" (Clark 1987: 40), examples can always be found at the margins of the categories concerned, examples can be found which might even be put in two places or which might move over the boundary between categories, as illustrated with the discussion over "interactive writing" above. Technological developments in relation to distance communication are breaking down the distinction between Speech and Writing, between Production and Interaction.

A more fundamental problem is that it is extremely difficult to avoid mixing different kinds of categories in the same set. In the list of activities in the grids presented earlier, there are in fact at least three kinds of categories, which are represented by different shapes in Figures 2.1 and 2.2 for Interaction and Production respectively.

Firstly, there are the activities related to the three macrofunctional uses of language discussed earlier (Transactional; Creative/Interpersonal; Evaluative). These are in the three large circles. The top circle in Figure 2.1 for Interaction (*Casual Conversation*) and in Figure 2.2 for Production (*Describing and Narrating Experience*—i.e. storytelling) are Swales' (1990) two pregeneric kinds of language use. The other two circles represent the Transactional Problem-solving requirements of social life. As befitting a pregenre, *Conversation* or "chat" can have a broad definition which would include most of the content in the other two circles (as intended in North 1992a) "to include chats as well as service contacts, therapy sessions as well asking for and getting the time of day, press conferences as well as exchanged whispers and sweet nothings" (Schegloff: 1972: 375). As Van Lier (1989: 500) points out, researchers are becoming increasingly aware that all

professional interaction is embedded in conversational interaction and "it is possible that a common core of general conversational proficiency underlies interaction in all specialised forms of spoken language use."

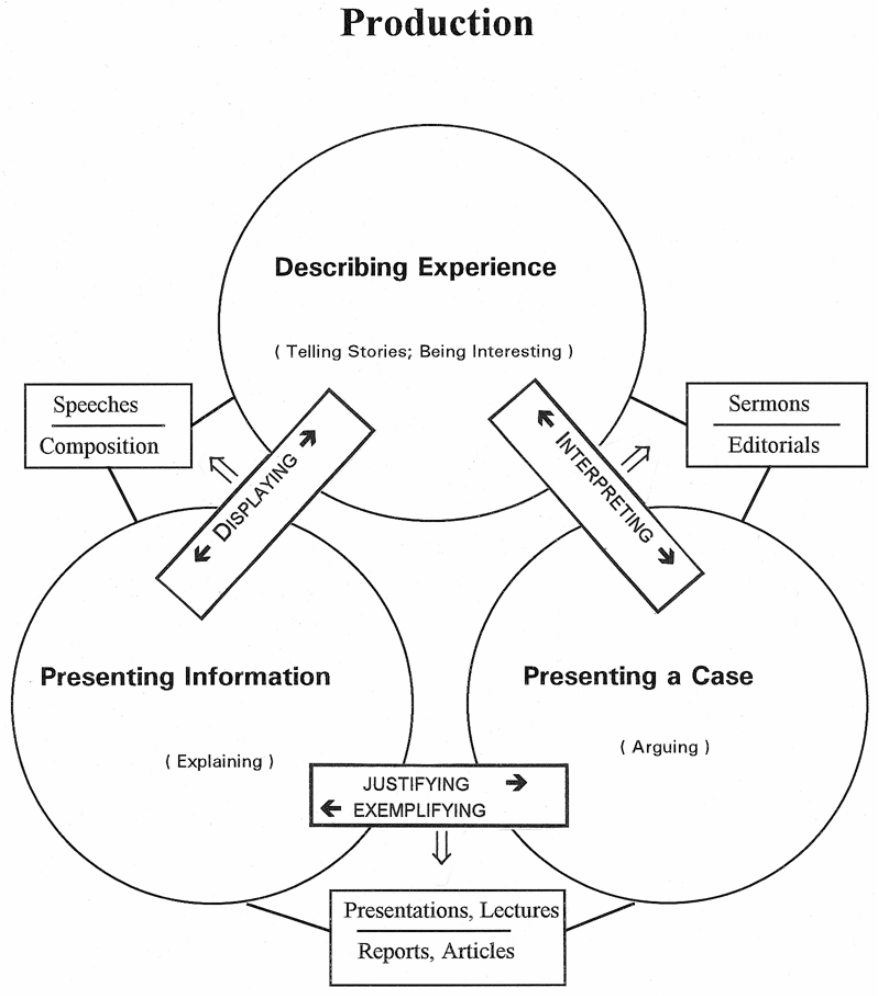
Figure 2.1: Boundaries of Categories of Spoken Interaction



Alternately, *Casual Conversation* can have a narrower definition as "a type of speech in which ties of union are created by a mere exchange of words" (Malinowski 1946: 315 cited in Ventola 1979: 278) of which the aim is "to create a friendly atmosphere, to establish contact, to forge new social relationships and maintain old ones" (Ventola 1979: 278). As mentioned above, Storytelling (*Describing and Narrating Experience*), the pre-generic

Production category (Swales 1990: 61), emerges from and is embedded in the turn-taking of this casual talk (Jefferson 1978: 220).

Figure 2.2: Boundaries of Categories for Spoken Production



But if *Conversation* can be defined to include the other circles in Figure 2.1, then these activities are not really completely separate. Moreover, a discourse which starts by focusing on one of them may very well slide into one of the others. A *Transaction,* buying a hamburger for example, may, through

complications and disagreement turn into a *Negotiation,* about the price/quality relationship for example, which may broaden into a general *Discussion*, particularly if other people get involved. That *Discussion* may lose focus (the seller has a built-in interest in its losing its focus) and, as people move off, may degenerate into a general the kind of "Isn't life terrible," "They really should……" i.e. a kind of *Casual* "*Café/Pub*" *Discussion*. The prime motivation of such discussion is phatic, social, to do with establishing and maintaining relationships. Conversely a *Casual Conversation* in which the other person keeps asking you for particular information starts to feel like a *Transaction*, and you begin to wonder what is going on—is he going to turn into a life insurance salesman?

Quite a lot of discourse at work, especially in corridors, could best be described in its own right as *Information Exchange*. Each person brings the other up-to-date with what is happening, which has a phatic, social purpose but also transactional one. As regards Production, if you are *Presenting a Case*, you may feel the need to exemplify by *Presenting some Information* to back yourself up; conversely in *Presenting Information* about a project, a product for example, you may feel a need to justify decisions made, choices taken and *Present a Case* that those choices were correct in the circumstances. Finally, a number of the particular genre (or groups of genre) which appear in scales of proficiency seem to be examples of formalised instances of such shifting discourse. *Interviews* are a classic example. An *Interview* is defined as "a scheduled encounter between unequal participants in hich one or more persons have vested rights to ask questions and organise the topic and talk" (Silverman 1976 cited in Ross and Berwick 1992: 169). As Berry (1981: 15) points out, there is a hierarchical relationship between participants, and one side *only* uses "frames" to organise the talk whether this be teacher-student, doctor/therapist-patient, examiner-candidate (or, one could add: employer-candidate). *Interviews* are formalised *Transactions,* predominantly question-and-answer *Information Exchanges* which masquerade as simulated *Conversation.* See Van Lier 1989; Berwick & Ross 1993 for a detailed discussion. Such formalised genre groups appear in the boxes outside the circles.

Advantages of a Communicative Activities Approach

As has been discussed above, there are inherent problems for any descriptive system in the way in which sub-categories relate to each other. However, it could be argued that a set of categories like those proposed is more capable of accommodating fuzzy boundaries and category shifts than is the

traditional division into the four skills and that a scale organised in such categories could therefore exert a positive influence or "washback" (Morrow 1986). In particular, such an approach might offer the following advantages:

- An organisation in terms of transparent activities in specific contexts might facilitate the recording and profiling of the "slices of life" (Selinker and Douglas 1985 cited by Douglas 1988: 257) which make up the language learner's experience.
- McNamara (1990, cited in Elder 1993: 248–9) finds that non-specialists focus on task success, so it might be an advantage to describe achievement in terms of the way people actually use the language. Two of the earliest scales of proficiency, ELTDU (English Language Development Unit) 1976 and IBM France (in Trim 1978), employ these kinds of categories rather than the four skills, and the categories being adopted by ALTE (Association of Language Testers in Europe) do the same.
- A set of categories which highlight the fact that the interpersonal perspective *and* sustained self expression are central by *Waystage* might help to counter-balance the pervasive transmission metaphor, which sees language "solely as the means by which information is shunted from one person to another…like sums of money or bags of oranges" (Smith 1985: 195).
- The distinction Reception, Interaction, Production might facilitate the integration of a broader concept of Strategic Competence. Cross-referencing Reception, Interaction, Production with Planning, Execution, Monitoring and Repair would produce a richer activity-oriented framework for describing Communication Strategies.
- A move away from the four skills for a new framework might also facilitate a general development beyond Lado's (1961) matrix of four skills x grammatical structure, vocabulary, phonology / graphology (still the norm for a great many teachers) and help promote communicative criteria for quality of performance.
- Taking relatively concrete types of use (tending towards supra-genres/speech events rather than abstract skills or functions) may help facilitate the provision of more concrete descriptors.

With the exception of the last point listed above, these are hypotheses which it is outside the scope of this study to investigate since they presuppose the existence of the instrument which it was the aim of this study to produce. Such a list suggests factors which might be taken into consideration in a future evaluation project aimed to collect further validity information. In that respect, however, one needs to bear in mind Wall and Alderson's (1993: 65–68) caution that an innovative instrument cannot by itself exert a positive impact; successful innovation is dependent on a range of factors and the instrument is only one of them. The most that can be said is that the pre-testing of categories and (where necessary) reformulation of descriptors with informants representative of the intended users can help to reduce the "innovation gap" and so increase the chances of such washback.

A Unitary Competence: Holistic Approaches

The last two sections have discussed in some detail categories to describe qualitative aspects of competence/proficiency, and categories to describe communicative activity. Apart from the question of what categories to use, there is the issue of the extent to which the categories interrelate, the extent to which they are subsumed into a higher or "global" proficiency, and the extent to which users can actually keep them separate, if they are separate. The fact that categories are conceptual artefacts and that any one category merges into its neighbours was discussed both in relation to attempts to validate a model of competence empirically and in relation to the development of a set of categories to describe contexts of language use. The question of whether categories or components can be combined into one unitary competence/proficiency (the UCH: Unitary Competence Hypothesis: Oller 1976/83) preoccupied language testers at the end of the 1970s. Bachman summarises the current state of play as follows:

> "The unitary trait has been replaced, through both empirical research and theorising, by the view that language proficiency is multicomponential, consisting of a number of specific abilities as well as a general ability or set of general strategies or procedures." (Bachman 1991: 673)

So there is still a place for a more modest concept of general proficiency. But in any case:

> "In asking the question of how language skills are "organised" one is really asking whether all skills are attained together, at the same rates, or attained separately, at different rates." (Carroll J.B. 1983: 83)

Most programmes of instruction in second language learning attempt to teach all these interrelated parts. Thus, different aspects of language tend to be learned together if they are learned at all, and advancement in any aspect of language is generally accompanied by advancement in other aspects. The "general language proficiency factor" reflects overall degree of advancement in different language skills—a function of the way the language is taught, the attention and effort the learner devotes to the study of the language, and possibly (or probably) the rate at which the learner is able to absorb and master what is being taught" (Carroll J.B. 1983: 94). It is certainly noticeable that there appears to be a contrast in correlations achieved between, for example, linguistic knowledge and oral communicative interaction, depending on whether the tests are taken at the end of a stay in an "acquisition-rich" environment (e.g. stay in a country where the language is spoken, attending a language school) or an "acquisition poor" environment (e.g. 2 hours a week foreign language classroom). In the former case correlations tend to be high, in the latter case they tend to be low. (For discussion of this issue see Tucker 1974 cited in Canale & Swain 1980; Upshur and Palmer 1974 cited in Palmer 1983; Ingram 1985). In discussing the effect of what might be called learning exposure profiles of different groups with correlations found between test scores, Cziko (1984: 26–28) takes reading and speaking skills as an example: if there is a very high correlation between amounts of exposure to learning opportunities in the two skills one should not be surprised if test results on the two skills are also highly correlated. This does not imply that the two skills are related, or manifestations of a unitary competence. It only reflects the fact that, as Carroll said, people tend to make similar progress in areas in which they are exposed to a similar degree to learning opportunities.

This is one reason some scales of language proficiency, which are pragmatic instruments generally associated in some way with progress in relation to "instruction" over time (or perhaps we might prefer to say to a learning environment), contain holistic scales for skills like speaking, and even global scales subsuming all four skills to represent this overall degree of advancement. The view of "global proficiency" taken in such cases is thus simply the aggregate of all grades across a profile, an overview of

where a person is in the system, which as Carroll and Cziko point out is an artefact of the learning situation. Other views exist, however.

Spolsky agrees with Oller (1983: 353) that holistic and analytic views are complementary and that:

> "…to say that linguistic and communicative competence are divisible does not necessarily rule out the claim that there is a core of common knowledge of a language underlying the specific abilities of a speaker…. *a knowledge*….varying in size from individual to individual such that you can rank individuals on the extent of their knowledge." (Spolsky 1985: 185–6; 1989: 71)

What Spolsky makes clear is that there are three ways of looking at the question of proficiency: general, functional and structural, and that they do not mesh neatly. While there are obviously three levels in a hierarchy (the more structures, the more functions; the more functions, the greater the global proficiency) "there is no one-to-one mapping between levels" (Spolsky 1989: 79) and one needs to be clear within which model one is working (Spolsky 1985: 189). Nevertheless, as Davies puts it "…we can see on the grounds of common sense…that, at some level, there is a unitary language skill, the level at which the distinctions among the performance skills of speaking, writing etc. are unimportant" (Davies 1991: 139–40).

Breen, coming from a totally different direction (task-based and procedural syllabus design) says:

> "This type of syllabus, unlike the formal or functional syllabus, does not take the four skills as the important manifestation of a language user's capacities, but calls upon those abilities which underlie all language use and which the four skills reflect in an indirect way. …Designers of a task-based syllabus believe that…underlying competence is generative in the sense that it is the means by which the learner can cope with the unpredictable, be creative and adaptive, and transfer knowledge and capability across tasks in ways that mastery of a fixed repertoire of performance might not facilitate." (Breen 1987: 162)

The kind of "generative underlying competence" being referred to here by Breen would seem to have certain similarities with the concept of "ability for use" put forward by Widdowson (1989) and Skehan (1995a; 1995b). However, the approach taken by Widdowson and Skehan is capable of explaining the lack of progress in grammatical competence and what Skehan

described as the "undesirable fluency" (Skehan 1995b: 553) of Schmidt's Wes, with his exclusive reliance on relexicalised knowledge (Skehan 1995b: 545) with no sign of "*effective* fluency…achieved when previous restructuring becomes automatized or becomes a (correct) exemplar" (Skehan 1995b: 553). Breen's implicit belief that participation in meaningful communicative tasks alone leads to the acquisition of linguistic forms (Interaction Hypothesis: See e.g. Hatch 1978; Givon 1979), on the other hand, is not supported by research for at least some types of learners (See e.g. Sato 1988; Nobuyoshi and Ellis 1993) who appear to require form-focus activities (Lightbrown and Spada 1990; Carroll, Swain and Roberg 1992) through the manipulation of pre and post tasks in relation to the particular activity (Foster and Skehan 1994; Skehan 1995b) in order to successfully generate more complex, more accurate language.

It is possible to combine all three views described above (Carroll, Spolsky, Breen) in an approach to "global proficiency." The Eurocentres Scale of Language Proficiency (used in classic "acquisition-rich" environments) has an overall or global scale as well as scales for the four skills, but the grade on the former is not an aggregate of the grades on the latter. Rather it is an average of grades on the scale awarded from (a) a test of knowledge of the language system: Rasch model item bank, cloze or c-test, depending on the language (what Spolsky is talking about) and (b) an extended communicative task involving group interaction (what Breen is talking about). In a controlled experiment at Eurocentres London, Lee Green, in summer 1992, the correlations between scores on tests from the itembank (Jones 1993) drawn mainly from the language specifications to the scale and the grades awarded through an assessment of oral interaction in small-group activities (North 1986; 1991; 1993b) was 0.87 (n=160). Such a level of correlation supports the view that a global scale can be appropriate for acquisition-rich learning environments which provide a balance between task-based and form-focused teaching. This by no means implies that one would be appropriate for those classic acquisition-poor environments, school foreign language classrooms.

In relation to the assessment of speaking and of writing a useful distinction has been made between unitary, global or *holistic* assessment scales, which follow Davies' argument above, and componential, category-based or *analytic* scales (Shohamy 1988b: 173). Holistic scales define global performance (e.g. "Speaking") whereas analytic scales focus on specific aspects or qualities of performance. These qualities may represent a theoretical model

of underlying competence—Clark's "hypothetical constructs of constructs" e.g. linguistic competence, socio-linguistic competence, discourse competence, strategic competence (Canale and Swain 1980). Alternatively they may reflect an operational model of supposedly more observable aspects like grammar, vocabulary, pronunciation, fluency (Shohamy 1981) or range, accuracy, delivery, interaction (North 1991). Many would agree with Davies that:

> "In testing as in teaching, there is a tension between the analytical on the one hand and the integrative on the other hand. It is likely that "progress" in language teaching consists of a dialectic between the two, indicated as a swing from the predominance of one emphasis to a predominance of the other." Davies (1990: 34)

As Davies implies, there are arguments for and against holistic and analytic approaches. A holistic approach encourages impression, intuition, and with it a tendency to:

> "…seize on a few salient or superficial points (errors of spelling, grammar or fact perhaps) and weigh those out of all proportion to the rest. On the other hand the analytic method by dealing with numerous isolated and possibly inessential points, may overlook certain general qualities that characterise the (essays) as a whole." (Cast 1939: 264 cited in Weir 1988: 70)

The analytic approach can perhaps be traced back to the work of Lado (1961) in which the four skills were generally divided into three categories phonology/orthography; vocabulary, structure. One has only to add fluency to arrive at the classic pre-communicative set of "factors" (e.g. FSI: Wilds 1975; Shohamy 1981). The problem with this approach, according to Stern (1992: 15) was that it assumed that language ability consisted simply of the ability to handle elements of the language system, whereas "language is not just the sum of its parts but the parts have to be mobilised and integrated together to carry out particular tasks in particular situations" (Ingram 1985: 229). On the other hand, "It was always recognised that the sum of the whole was greater than any one of the parts" (Davies 1978: 216 cited in Vollmer 1983: 7) and always "more or less implicitly assumed that each of the cells in the matrix (the four skills by the three factors) is related to an

underlying competence" (Vollmer and Sang 1983: 35; Carroll 1967a: 49 cited in Carroll 1983: 95).

Holistic scales interweave points relevant to different aspects of competence into one descriptor (Hofmann 1974: 5) making certain points salient at particular levels where this seems intuitively appropriate where they operate as "key parameters (*which*) help the rater to identify where the learner's proficiency falls on the scale" (Ingram 1984: 5). A holistic approach is sceptical whether competence can be "untwined" into components, or, by implication, described in a definitive fashion. It relies upon the fact that once people know the system, they use their intuitive judgement to recognise "a Level 2", and probably only refer to the written descriptions occasionally (Jones 1981: 105; Jones 1985: 77; Berkoff 1985: 96). It is this internalised concept, this common frame of reference arrived at through training, which is primary.

But this holistic approach, whilst pragmatic and popular, has been criticised on three grounds.

Firstly, the scale may say more about what the person can do in general than spell out "criteria for acceptability" in relation to qualities of any particular performance (Pollitt 1991). This relates to the purpose and orientation of the scale, discussed in Chapter 1.

Secondly, with a holistic approach to assessment, no one knows what the salient criteria of the *raters* are; when making judgements people may be thinking of quite different things. Some may have a more analytical, bottom-up, linguistic rating style, other a more holistic, top-down, communicative, rating style. Introducing analytic scales for different features of performance can help to provide a common metalanguage in which to negotiate grades, debate the nature of proficiency, give feedback to learners and even involve learners as (self)-raters.

Thirdly, with holistic scales which are developed too quickly, a co-occurrence of certain features may be assumed in a way which is counter-intuitive and not verified empirically by reference to the features of actual performances (Skehan 1984: 217; van Ek 1987: 24; Fulcher 1987, 1988). Even when the mix of "a multiplicity of behavioural criteria in one level" is done on the basis of experiential verification, it can be argued that the fact that these different features are "collapsed into one single dimension and subsumed under a number does not accord with current thinking about the multidimensional nature of language" (Brindley 1986: 56). In discussing the

Scottish Standard Grade and English GCSE exams, for example, Clark describes the descriptors of the quality of performance required as:

> "…at best gross averagings out of important individual variations, and at worst fictional distortions of reality. We must guard against giving them any ontological status, and treat them rather as the abstract artefacts that they are, however much impressive evidence may be collected as to the reliability with which trained teachers can allocate learners to the various grades or levels described." (Clark 1987: 46).

As Skehan, Van Ek, Brindley and Clark say, the problem with such holistic qualitative statements is that they appear to assume that learners develop in all aspects of communicative proficiency in roughly the same way at roughly the same rate. All four of them propose profiling on the components of competence, but avoid specifying which and how many components would be appropriate for that purpose.

However, a fact proponents of multidimensional, analytic scales sometimes overlook is that if such a model is to be used for assessment, then it needs to be a pragmatic, workable one. Human beings have a limit to the number of factors they can juggle in information-processing in short-term memory. The *maximum* number of categories (not chunked / aggregated into larger categories) is 7 plus or minus 2 (Miller 1956). It is perhaps not surprising that virtually all practical approaches use 3–6 categories. Keeping to a small number of categories means that some categories must be "subsumed" under others, "collapsed" assuming "co-occurrence." So opting for an analytic model does not remove this problem of aggregation, though it may reduce it. There is also the problem of "halo effect," which is the apparent incapacity of raters to keep the hypothetical constructs of competence separate, a tendency to give the same grade for the different categories, cross contamination:

> "One would expect, and indeed one gets, different performance on different dimensions (such that it is possible to get, say, 3 for appropriacy, and a 5 for content) and it is undesirable to add the scores on the separate dimensions together in order to arrive at some global assessment, because individual differences will be hidden in such a procedure: what is required is the reporting of some sort of profile. However, the question was raised of the independence of such dimensions, if not in reality, then at least in the ability of judges to rate independently. Cross contamination is quite

> likely, and only avoidable, if at all, by having different judges rate performances on different dimensions." (Alderson 1981: 61)

Halo effect is a complex problem which has at least partly to do with the tendency for raters to assign people to prototypes: the more "typical" the person, the greater the halo effect (Mount and Thompson 1987: 244).

Alderson's suggestion of multiple raters is not practical in most settings. Furthermore research both outside and within language learning has shown that rating all candidates on one category before moving onto the next does not solve the problem (Landy and Farr 1983: 149 citing work 1956–68; Yorozuya and Oller 1980). And as Alderson reports in relation to IELTS revision, asking raters to concentrate on different criteria in different phases of an interview was abandoned because raters found it "simply too complicated" (Alderson 1991a: 79).

Part of the problem is that, as actually assumed in "holistic" scales, many of these categories do in fact co-occur in performances of a certain type of people in a certain context, so part of the "halo" is in fact "true halo" and not an erroneous "illusory halo" at all (Cooper 1981). Halo says that the rater *didn't* discriminate, not that he/she *couldn't* (Murphy and Cleveland 1991: 22). These problems relate to assessment rather than description. There may be perfectly good reasons for wanting this diagnostic detail, but assessing *through* detail may not be the way to get it, since many people have difficulty seeing the wood for the trees.

One possible solution might be to separate profile assessment (describing) from level assessment (rating) (Hulstijn 1985: 280). This was in fact the practice of the FSI assessors and is also applied in the ILR interview and in the Ontario OTESL Oral Interaction Test. For the ILR the assessor gives a grade for different factors after assigning a holistic scale band; in the Ontario test, the assessor first assigns one holistic band, but makes notes on particular qualitative aspects of the performance on each task. In both cases the motivation is diagnosis not assessment. Such a separation of profiling and assessment could perhaps also be achieved in a different way during teaching programmes by blending (a) continuous assessment of coursework in discrete "series tasks" (Harrison 1982b) using simple mastery criteria with (b) assessment at particular points (e.g. at half and end of term) of qualitative aspects of proficiency (category assessment: Harrison 1982b), reporting a holistic overall level for the skill concerned, but making notes on the specific categories.

Towards Balanced Scale Categories

The fact that, despite considerable consensus, no universal, validated, theoretical model of either communicative competence or of communicative activities exist or is likely to exist for some considerable time leaves one, as Clark (1986: 62) stated, with a pragmatic choice. Part of that choice is a decision between *theoretical constructs of applied linguists* like "socio-linguistic competence" (Skehan 1984: 209) and *operational models* like the criteria for degrees of skill in more observable aspects of performance suggested by B.J. Carroll (1980) and Morrow, and used in suites of communicative examinations (e.g. University of Cambridge/Royal Society of Arts 1990). Such operational approaches take account of theories of underlying competences, but regroup them in different ways in order to focus on features which are more observable, or which it is felt should be highlighted with regard to this particular task. However there is an apparent danger that, unless guided by a theoretical model, such pragmatic decisions can lead to the choice of somewhat impoverished rating criteria.

On the other hand, people can generally only make progress from where they are. Attempts at innovation often fail because they do not connect with the state of the art out in the field. The gap between current, proven (even if imperfect) practice and proposed innovation is often too great for the practitioners to bridge. White (1988: 140) cites Rogers (1983: 15–16) that the rate of adoption of an innovation is directly related to the receivers' perception of its attributes. Rogers proposes six factors related to adoption rate: Relative Advantage; Compatibility; Trialability; Observability; Degree of interconnectedness in a social system; Complexity. In relation to the selection of categories, Compatibility and Complexity are the most relevant, perception of complexity being negatively correlated to adoption. In other words, if one tries to present or introduce an apparently complex descriptive model which is apparently incompatible with current practice, its chances of adoption will be slight. The perceived Relative Advantage would have to be overwhelming, the procedures associated with it would need to be highly Observable and Trialable (White gives video and groupwork as examples of highly observable trialable innovation), and the degree of interconnection in a network of project groups would have to be very extensive indeed if such a complex approach, requiring so much reorganisation of current procedures (due to incompatibility) were to succeed.

The lessons for the development of categories and descriptors from the point Rogers and White are making would seem to be:

1. Keep in mind theoretical models, but start from where the field is. The "field" in this case being: (a) existing scales of language proficiency and examinations (b) practising teachers representing the range of qualification and experience concerned.
2. Ensure new categories can be seen as either developments from old categories or as offering a clear relative advantage (e.g. for strategies).
3. Ensure that old comfortable categories (e.g. Listening, Speaking; Fluency, Accuracy, Pronunciation) can be easily located in the new wider model.
4. Ensure that practitioners can distinguish between the categories, feel comfortable with them and feel they offer a Relative Advantage.
5. Ensure that practitioners can understand the descriptors, agree with what they say and find them useful and relevant.

In an attempt to take account of these points, and arrive at a pool of descriptors which would go some way to bridging the gap between theoretical and operational models, the first phases of this study (December 1993 to April 1994) consisted of an interactive process involving the development of a set of descriptive categories in the authoring team for the Council of Europe Framework; the analysis of existing scales of language proficiency; a series of practical workshops with teachers, followed by a refinement of the categories.

Firstly, the content of the 30 odd proficiency scales used as sources were edited into descriptors which were grouped under suitable categories, which were themselves grouped under higher order categories for Communicative Competence and for Communicative Activities simultaneously being developed in the Council of Europe authoring team. A framework for Strategy Use was also developed, and a number of descriptors for it written.

Secondly, in a development of a technique used by Pollitt and Murray (1993) based on Kelly's theory of personal constructs (Kelly 1955 cited by Pollitt and Murray 1993), teachers' discussions in a series of small workshops in which they compared two performances on video were recorded and analysed in order to identify the categories used by the teachers to make comparative judgements at different levels and check that these categories were described in the pool of descriptors.

Thirdly, in a development of a sorting technique used by Smith and Kendall (1963), in the same series of workshops teachers sorted descriptors

into categories to check that the categories being used made sense to teachers, and that the descriptors described what they were intended to describe. Following Smith and Kendall, descriptors which were frequently misplaced were discarded or reworded and categories which appeared ambiguous were discarded or revised. The process is described in more detail in Chapter 4.

The result was the production of the set of categories in Table 2.15.

Table 2.15: Categories for Descriptors used in the Study

Aspects of Communicative Language Proficiency	**Strategy Use**	**Communicative Activities**
Linguistic:	**Reception:**	**Interactive Listening:**
General range	Framing	Understanding a NS
Morpho-syntactic range	Inferring	Following discussions between native-speakers
Vocabulary range	**Interaction:**	**Interaction:**
Grammatical accuracy	Turntaking	Conversation
Vocabulary control	Cooperating	Discussion
Phonological control	Getting clarification	Formal meetings
Pragmatic:	Asking for help	Negotiating
Fluency	Repairing	Service encounters
Flexibility		Information exchanges
Coherence and cohesion		Interviews and interviewing
Message precision		Personal letters
Thematic development		Formal letters
		Completing forms
Socio-cultural		Writing notes and messages
	Production:	**Production:**
Independence:	Planning, rehearsing	Describing and narrating
Need for speech adjustment	Compensating	Putting a case
Need for clarification	Monitoring	Presentations
Need for help		Writing articles and reports

Communicative Activities have not been fully elaborated in this version, since this study focused on Interaction and Spoken Production. Some descriptors for Listening in Interaction and for Writing were included in the descriptor pool, but descriptors with a purely Receptive focus (e.g. on Listening to Public Announcements) were excluded, since they were to be the subject of the follow-up survey in 1995.

In relation to Linguistic Competence, descriptors were provided for both Range and Accuracy rather than just Grammar and Vocabulary in order to focus attention on breadth and complexity of language resources as well as the correctness of their use. Descriptors avoided references to specific language forms.

In relation to Pragmatic Competence, the reason for placing Fluency there was discussed earlier. Flexibility concerns the extent to which the user can adapt his/her available linguistic resources to the context in question.

The English National Curriculum for example has a number of descriptors of this type, of which two were merged with a similar statement from Wilkins (1978) to give: *Can adapt well rehearsed memorised simple phrases to particular circumstances through limited lexical substitution*. A higher level descriptor, in a linguistic as well as proficiency sense of "level" was merged from elements in Trim's (1978) 4th level ("Adequate response normally encountered") and ACTFL Advanced to give *Can adjust to the changes of direction, style and emphasis normally found in conversation.* Flexibility does show some relationship to Range (Linguistic) on the one hand and to Planning (Strategic) on the other, especially with descriptors like the first one. It was retained as a separate category under Pragmatic (i.e. Language Use) because it talks about what the learner does with resources rather than the resources themselves, and because in both sorting tasks and recorded discussions, teachers related it closely to Fluency. Coherence & Cohesion and Thematic Development are included under Pragmatic as aspects of discourse competence, Message Precision as being a manifestation of speaker meaning.

The difficulty of describing and scaling Socio-cultural Competence was also discussed. Much of what is put under Socio-cultural Competence concerns knowledge of the world, which Bachman (1990a: 85) separates from knowledge of language, and could therefore be expected to be on a separate dimension to language proficiency. Also, descriptors for Socio-cultural Competence which try to summarise ability at different levels tend to be rather vague and open to differing interpretation. Both these factors pose problems for scaling in relation to a proficiency model.

Independence is a term used to cover relational factors involving accommodation of the lack of proficiency by the learner on the part of the interlocutor: the need for a (sympathetic) interlocutor who adjusts his/her own speech to help the language user follow it, and the need the language user has to ask for clarification.

In relation to Strategy Use, Getting Clarification is a positive strategy to deal with the lack of Independence mentioned above; Compensating and Asking for Help are positive strategies to cope with lack of Linguistic Range. Repairing and Monitoring are positive strategies primarily to pick up and correct grammatical and vocabulary errors and to re-establish communication when it breaks down. Turntaking in the sense of taking the turn (discourse challenge: Burton 1980) and Cooperating Strategies are interactive discourse strategies and could be regarded as part of discourse and hence Pragmatic Competence, rather than as examples of Strategy Use. Van Ek (1987) faced with this dilemma, puts them in both places. Here they have been included under Strategies for two reasons. Firstly, because they take place only in Interaction, and are related to the other Interaction Strategies: Asking for Help and Getting Clarification both involve "discourse challenge." Cooperating to manage discourse involves taking the turn to invite others in, and repeating what others have said or relating what you say to what they have said as a way of getting clarification that they meant what you thought they meant. Secondly, Pragmatic Competence/Language Use is here confined to the speaker/writer's ability to articulate their meaning, to express what they want to say. It can be applied to both speech and writing, to both Interaction and Production, as can Linguistic and Socio-cultural Competence. Taking the turn and Cooperating, on the other hand, are activities through which joint meaning is constructed interactively.

The scheme presented for categories for Communicative Activity and the fluidity of some of the boundaries between categories was discussed. A fundamental aim of this type of presentation is to separate what one teacher in the survey (my wife) described as "the ping-pong skill" (Interaction: short turns: Brown et al 1984) from coherent language Production (long turns: Brown et al 1984). Several secondary teachers in fact commented that the so-called "communicative" emphasis in beginner and elementary textbooks written for the Swiss school market overemphasised transactional "ping-pong" exchanges, box-ticking and form-filling and did little to develop creative interpersonal language use either in Interaction (conversation skills)

or in Production (describing and narrating). A distinction about which teachers were enthusiastic was initially made between Casual Conversation, Casual Discussion, Goal-oriented Collaborative Discussion and Meetings. Casual Discussion was discussed in terms of the kind of Café/Pub discussion which range widely but is not seriously seeking to find solutions—a kind of thematic conversation with airing of views. Unfortunately it proved impossible to establish *any* firm boundary between Casual Discussion and Goal-oriented Collaborative Discussion. Also many good descriptors about Discussion belonged in neither, so the decision was reluctantly taken to revert to a more conventional distinction between Conversation (emphasis interpersonal; expectation two people) and Discussion (emphasis ideational; expectation more than two people). Formal Discussion is discussion chaired to an agenda, normally but not exclusively associated with meetings.

Summary on Description Issues

This chapter has argued that whilst there is a considerable consensus between models of second language proficiency, there is unlikely to be an empirically validated generalisable description of proficiency in the foreseeable future. Secondly, although a number of moves are taking place towards a reorganisation of the four skills into something closer to the way language is used, no consensus model has emerged. Decisions on categories have therefore been made informed by, but not dictated to, by theory, and with some validation for the context in question through pre-testing in the series of workshops described in Chapter 4. The set of categories tries to take account of the perspectives and needs of both insiders (teachers, assessors) and outsiders (potentially learners, employers, parents) in covering activities in which language is used, strategies used to achieve goals, and factors in the quality of language demonstrated.

A common framework scale based on such a set of categories could have validity only in relation to the (admittedly multi-lingual, multi-sector) context in which it is developed. There are no universal solutions (Widdowson 1990: 23), no panaceas, no ideal language curriculum to suit all circumstances (Clark 1985: 343–4), no definitive set of categories. Boundaries between the categories, which are all conceptual artefacts, are not watertight.

Although descriptors on existing scales used as sources have been "deconstructed" into the categories employed, the process can only be taken so far without reducing definitions to the status of a list. Descriptors for one

category sometimes contain a hint, or more than a hint of another category. Such descriptors were often split into two or more variants, but when the original appeared to be saying something coherent and concise, and was consistently judged to be clear and useful by teachers in the workshops—especially if they put it consistently into the "right" category—it has been retained. In other words, whilst there has been an effort to develop a coherent, comprehensive set of categories, and to group, edit and write descriptors for those categories, descriptors which were clearly good descriptors have not necessarily been sacrificed to the boundaries between categories, since it is quite probable that users of the resulting calibrated bank of descriptors which go to make up the scale may in any case refine, alter or elaborate the categories in order to meet local needs, new purposes (e.g. self-assessment, oral test criteria), new contexts (e.g. university level) or new pedagogic cultures (French and German as foreign languages).

Apart from the pragmatic nature of the decisions taken on categories, and the relative nature of any validity which could be claimed for them on the basis of the analysis described in Chapter 6 there is the additional fact that categories may be perfectly good categories conceptually, but not meet the requirements of a measurement scale: that people can be consistently judged to be better or worse at this category. Alternatively, it may be the case that the category has integrity in its own right, and could be used to construct a measurement scale, but just does not belong in the same dimension as the other categories included in the scale of language proficiency, and hence cannot be scaled with them. On the other hand, the category could be fine, but the descriptors used to define it could fail to meet the requirements for a valid scale (e.g. be too vague). Yet again, the category and descriptor could both be fine, but raters might be incapable of making separate judgements about different things in order to rate a learner in relation to the specific criterion descriptor. Or they might be able to do so if they were not confused by having to make too many decisions between steps on the rating scale. Finally the categories, descriptors, ability to separate things, and number of rating scale steps may all be fine, but people may do it so subjectively that the results are inconsistent and unreliable.

These are all measurement issues, and are discussed among other measurement issues in Chapter 3.

3 Measurement

The reason that both naive native speakers and language testing raters tend towards a holistic strategy is that evaluation is inevitably a comparative process, and hence:

> "Evaluations fall along the same continuum, allowing for meaningful comparisons across persons, tasks, persons and the like. The universal nature of this general evaluation has been noted in several separate research literatures. …The existence of a single and universal dimension implies that the consistent scaling of performance is at least theoretically possible." (Murphy and Cleveland 1992: 119–120)

When Scriven, inventor of the term formative evaluation (Scriven 1967; cited in Widdowson 1990: 51), was asked if evaluation was always comparative, he is reported to have replied "No, only g*ood* evaluation is comparative" (cited in Glass 1978: 259). This is arguing against the simplistic division into master / non-master; competent / incompetent which has dominated thinking in criterion-referenced assessment: "knowing and being able to are not absolute, all-or-nothing attributes" (Trim 1978: 51).

Criterion-referenced Assessment

An intuitive definition of criterion-referencing is that students' performances are judged in relation to a defined standard and not in relation to their peers. The standard defines where the learner is on the continuum of learning. Thus as he/she learns he/she progresses through standards and makes visible progress even if his/her position in the class (e.g. 7th) remains unchanged. This is in opposition to norm-referencing, which defines a student's place in relation to his/her peers: rank in class (on school reports) or in school cohort across the country. In a traditional British examination context, a specific proportion of the candidates, say for the sake of argument 34%, would be doomed to failure, since a primary function of the education system was to sort people out into groups for different types of

education for different kinds of lives. The system could go to extreme lengths to preserve this social role and prevent mere successful learning "distorting" examination results. At Aston University, an early application of computer-assisted learning in a department of the engineering faculty led to such outstanding results that the examination board first accused the students of cheating and then set a more difficult examination in order to preserve "fairness" (Croxton and Martin 1970 cited in Romiszowski 1981: 36)!

Scales of language proficiency and profiling systems are generally considered by their developers to be criterion-referenced. Learners are assessed not in relation to each other, but in relation to defined stages on a continuum, which proponents of scales of language proficiency would identify with the continuum defined by Glaser, the founder of criterion-referenced assessment when he talked of:

> "a continuum of knowledge ranging from no proficiency at all to perfect performance.… Along such a continuum of attainment, a student's score on a criterion-referenced measure provides explicit information as to what the individual can and cannot do. Criterion-referenced measures indicate the content of the behaviour repertory, and the correspondence between what an individual does and the underlying continuum of achievement. Measures which assess student achievement in terms of a criterion standard thus provide information as to the degree of competence attained by a particular student which is independent of reference to the performance of others." (Glaser 1963: 519–20)

There are two basic perspectives on criterion-referencing, which have emphasised different aspects of Glaser's statement, *the state model:* dividing learners into masters and non-masters, (Meskausas 1976: 142) focusing on the *achievement* of a criterion standard, and *the continuum model:* situating learners on a continuum of ability (Meskausas 1976: 134) focusing on *proficiency* in relation to "a continuum of knowledge ranging from no proficiency at all to perfect performance."

Bachman discusses the way in which the most common interpretation of criterion-referenced assessment—that "test scores are reported and interpreted with reference to a specific content domain or criterion level of performance" (Bachman 1990a: 8) lost the continuum aspect because of educators' emphasis on achievement tests within a particular scholastic system (Bachman 1990a: 242–3) rather than what students could do in the real world.

Skehan identifies four applications of criterion referencing to language testing (Skehan 1989b: 6). His distinction could be glossed as follows:

1. *External Scale:* Tests referenced to an external scale covering the continuum of proficiency. Examples: the TOEIC which is referenced to the ILR scale, or Eurocentres Itembanker, which is referenced to the Eurocentres scale. Such tests do not need to "look" criterion-referenced.
2. *Mastery:* Emphasis on a cut-off score, a criterion level of performance. The classic US mastery learning approach which can be applied to language testing, complete with a revised approach to classical testing statistics (Hudson and Lynch 1984; Hudson 1989; 1993).
3. *Domain:* Domain tests assess an area which has been specified in detail. In mainstream US measurement circles this is sometimes interpreted to mean so tightly specified that it should ideally boil down to an algorithm which could produce all possible test questions (Baker 1988). More holistic examples would be the RSA tests for EFL and modern languages (University of Cambridge/RSA 1990; Royal Society of Arts Examinations Board 1989) which have detailed specifications of both test tasks and degrees of skill required in performance
4. *Levels:* "But there is a fourth sense of CRM which is even more difficult to achieve and which has eluded language testers so far. This is that the proficiency levels which are the basis of criterion referencing are linked in some cumulative way to a course of development. This would allow each intermediate step simultaneously to have a proficiency, real world dimension, and also a relationship with other stages of development. Each stage would be measured by a cut-score approach based on items which sample a defined domain, but the stages themselves would be cumulatively meaningful" (Skehan 1989b: 6).

This fourth sense of CRM is the aim of integrated systems of levels, assessment and certification, like, for example, the Eurocentres framework. There are two ways of getting to Skehan's fourth, ambitious interpretation of CR: from a state (or mastery) model and from a continuum (or scalar) model.

The Scottish GLAFLL project (Graded Levels of Achievement in Foreign Language Learning) offers an example of the former. The philosophy of the UK graded objectives movement was criterion referenced (Page 1983). Baumgart (1986) suggests that graded objectives schemes follow a "meta-state" model, which organises a series of learning tasks in an implicational hierarchy—master this first—with a summative examination made up of tightly defined modules on skill areas to report a profile. He suggests that:

> "If it is necessary to divide a whole curriculum into a number of ordered levels of mastery (as in GOML schemes or music examinations) the meta-state model described above would seem to be flexible and more pedagogically sound than a continuum model. Thus in order to show progress from one level to the next, a student might be expected to show competence in each of a number of features, but the features themselves might only be loosely associated or even independent. Schematically, one might think of a student progressing through a series of ordered plateaux (levels). But to progress from any one plateau to the next, the student might be expected to climb (or show competence in) each of several towers (features) with ordered tasks within them." (Baumgart 1986: 53)

Within the graded objectives movement itself, a distinction has been drawn between the English schemes (GOML: Graded Objectives in Modern Languages) on the one hand (state or meta-state model) and the Lothian, Scottish, scheme (GLAFLL: Graded Levels of Achievement in Foreign Language Learning) on the other. The Lothian approach was directly inspired by *Threshold Level* as well as the pioneer GOML schemes in York and Oxfordshire (Page 1983: 296, Clark 1987: 140–145). However:

> "GLAFLL syllabi are not preoccupied with functional-notional objectives which in assessment jargon are behavioural objectives but rather with a range of skills and knowledge involved in communicative performance which might be best identified as process objectives. GLAFLL is thus based on the concept of graded levels of linguistic performance which are themselves broken down into overlapping differentiated levels of achievement." (Martin 1985: 11)

This GLAFLL interpretation is more like a continuum model, although there are no holistic descriptors of outcomes at the different stages, no actual scale of proficiency. The room for manoeuvre allowed by a more holistic approach "to do with learners building up a language resource" rather

than with "language being broken down into digestible chunks for learners to master" leads Martin to label the GOML model "mechanistic" and the GLAFLL approach "humanistic." This distinction seems somewhat exaggerated, but when the categories used in assessment and profiling become too small, and criterion-referencing is interpreted as determining mastery / non-mastery of lists of tasks, functions, topics structures etc. there is a danger that the whole process can become bureaucratic and can turn the teacher into a "Gradgrind, more concerned with measuring out pupil performance into various little pots of his own making than with getting on with teaching" (Clark 1986: 62). The GLAFLL approach avoids this pitfall by profiling at macro-level of stages of attainment rather than at the micro-level of all the content elements within them.

Eurocentres offers an example of the opposite way to develop a set of criterion levels, from a continuum model. The Eurocentres approach is based on task analysis. The descriptors in the different sub-scales of the scale were analysed to produce a list of communicative tasks which are mentioned or implied. An analysis was made of the functions necessary to perform those tasks. Sample exponents to express those functions which appear appropriate for the level concerned are selected, and examples given. The result is a set of *content specifications* of tasks, functions and structures very much following the method of analysis used in the "*Threshold Level*" but in this case related to scale levels. Unlike in a state/mastery model (and more like the GLAFLL scheme) these specifications are advisory, in the form of a guideline for the teacher, who designs their own scheme of work in negotiation with the class. Assessment takes place from both a state angle, by continuous assessment of achievement, and from a continuum angle, through assessment of proficiency in communicative performance in relation to defined degrees of skill required for each level. These degrees of skill form *outcomes specifications,* which are presented in the form of rating scales and grids.

The Royal Society of Arts Modern Language Examinations, the UK Language Lead Body National Language Standards and the Leeds Language Project (Killick 1992) all also represent the merging of approaches from specification of objectives (like *Threshold Level*) and associated state or meta-state assessment models (like GOML) which are *content specifications* on the one hand, and a continuum model represented by a series of ascending performance levels, a scale of proficiency, *outcomes specifications* on the other hand.

The fact that all the examples given of Skehan's fourth category of criterion-referencing are British is not an entire coincidence. It is a reflection of the impact of the notional/functional approach and *Threshold Level*. In discussing the communicative approach, Stern draws a distinction between the North American "P" (psychological/pedagogical) approach to communicative language teaching and the European "L" (linguistic) approach (Stern 1981: 134–140). The tradition of linguistic analysis of language use in context leads to the definition of content specifications as well as outcome specifications.

Criterion-referenced Assessment and Scales of Language Proficiency

Even though most authors of scales of language proficiency would consider that the scales provide a continuum model of criterion-referenced assessment, one should note that a number of specific points have been made as to why at least some scales of proficiency are *not* true criterion referencing as defined in the literature (Skehan 1984; 1989b; Brindley 1991).

Relative Levels. They represent a continuum model yet lack an absolute scale from zero to perfection, being scales of levels which exist only relative to each other (Bachman 1990a: 343–4). It is argued later that this point is misguided. Firstly, as Bachman himself agrees (1990a: 26; 36; 44–5) existing scales with this characteristic would classify as ordinal scales, secondly, it can be argued that absolute zero and perfection points are not essential to either criterion-referencing or to valid measurement scales.

Relative Wording. Most scales, especially those which are not based on functional or activity categories, employ relative, normative wording:

- relational concepts given a pseudo absolute character e.g. intelligible, comprehensible (Trim 1978: 6; Brindley 1986: 19);
- norm-referenced terms: e.g. "generally can get by without serious breakdowns" or "better than an absolute beginner" (both from Carroll, B.J. 1980: 134 and cited by Skehan 1984: 217);
- qualifiers to distinguish between levels e.g. switching "limited" to "moderate" (North 1992a: 167). This is very difficult to avoid, especially when describing aspects of competence (Alderson 1991a: 81–2).

Failure to offer a Yes/ No criterion. This vagueness of the wording of descriptors often means that a single descriptor can only be interpreted by contrasting it to the descriptors for the level above and below on the scale. As Skehan points out, this is not the spirit of criterion-referencing:

> "…the "criterion" that is referred to is different in meaning from that referred to in a true criterion referencing. It is not linked to any specific performance which would allow a definite "yes/no" judgement on whether a particular task can be accomplished. Instead, the criterion is usually made up of generalised criteria-linked behaviours which attempt to relate test performance to "objective" or "external" criteria which can be applied to real life tasks." (Skehan 1984: 217)

Circular Argument. There is often a certain circularity in the argument: a text is Level 3 because people who are Level 3 can read it; people are Level 3 because they can read texts which are Level 3. At the end of the day Level 3 is what we decide it is, so it is a standard, but is that a criterion? Is it a criterion if it is specified in detail? Is it still a criterion if that specification in question turns out to bear little relationship to (a) actual as opposed to supposed difficulty of texts (Allen et al 1988) and what actually happens when people read texts? Dandonoli (1987) discusses these problems in relation to developing computer-adaptive reading tests for ACTFL. The same argument about circularity applies to speaking (or writing) though one can provide samples of performance. Can the criterion description be valid if there are no samples? Presumably it loses validity if samples or research indicate that what actually happens is different from what is described in the descriptor (Fulcher 1987; 1988). Does the criterion descriptor become (more) valid if you try and take account of such features in the development (Ingram and Clapham 1988; Alderson 1991a)? What if there is a test external to the subjective circularity which acts as a criterion, as the Eurocentres Rasch model "Itembanker" does?

Some consider the circularity an indication of the coherence of the system, others consider it shows that the definitions are "a self-contained closed system" (Lantolf and Frawley 1988: 186) "lovely symmetrical definitions, but that is all they are" (Lantolf and Frawley 1992: 35).

Criterion-referenced and Standards-oriented Assessment

The argument about circularity is at least partly caused by the difficulty of distinguishing between criteria, standards and norms. The distinction be-

tween norm- and criterion-referencing itself is not really as absolute as people sometimes appear to suggest.

There can be ethical objections to criterion referencing as well as norm-referencing, especially if the former is interpreted as dividing people into two groups: masters and non-masters:

> "The kinds of tests used in schools are only indirectly related to common sense notions of what it means to be a competent language user. When notions like *the ability to extract meaning* become operationalised as scores on, for example, reading tests, a child who fails is then labelled as one who is *unable to extract meaning*." (Martin-Jones & Romaine 1986: 29)

This might be okay if one could be sure one was setting a fair standard, but "if the literature on standard setting is conclusive on any point, it is the difficulty of setting defensible standards on competency tests" (Jaeger 1989: 491). It is a very subjective process, and teachers and examiners have in practice great problems predicting the difficulty of test items, or even what they are testing (Alderson 1990a 1990b,1995).

For every criterion-referenced test, there must be a population for whom the tests could be norm-referenced (Davies 1990: 18–19). Common sense indicates that that norm should not be too far away from the norm (average mark) of the group taking the test. It is in fact:

> "…extremely difficult to imagine a criterion-referenced assessment that is totally independent of norm levels. If a criterion point is to be useful, it must obviously distinguish between individuals, so that the possibility exists that some of them will reach it and others will not. We can know whether this is the case only by collecting data on how many individual students do so, thus estimating a norm. Since in fact "cut off" points are a choice, if not actually arbitrary, this "norming" information will typically be used in determining them, at least initially." (Nuttall and Goldstein 1986: 186)

Levels on a scale of proficiency really operate as standards. If the levels have a pragmatic origin in terms of likely stages of attainment, they are also likely to be derived from "norms" for those stages in the sense of what it is reasonable to expect. In trialling the system and seeing if the descriptors do describe what was intended (whether the learners "get there"), the same kind of "norming" information may be used to adjust the cut-off. It is just

that in this case the cut-off (standard) may be in a formulation of expectations rather than a numerical score.

For these reasons, evaluation in relation to a scale of proficiency is sometimes referred to as *standards-oriented assessment* (Sadler 1987; Gipps 1994).

The Development of Behaviourally-based Assessment Scales

In relation to the question of whether scales of proficiency are criterion-referenced assessment, but also in relation to other issues such as the gain to be achieved from precision in the definition of what is expected at different levels of proficiency, and the question of how to assign such definitions to different levels, it is instructive to look at the development of defined behavioural criteria in rating scales in work evaluation. This literature offers many parallels with scales of language proficiency in the ways people have tried to give meaning to numbers, to describe the features being assessed.

Before the arrival of numerical rating scales in the aftermath of the First World War, all one had was weighted marks for undefined characteristics or dimensions (See Borman 1986: 102 for examples). Classic numerical rating scales had just a label (undefined) for the dimension and boxes to tick, as indeed one still sees in many foreign language examinations, for example the Kleines Deutsches Sprachdiplom from the Goethe Institute, or the Diplôme d'Etudes en Langue Française. A slight elaboration was a short definition of the "dimension," "trait" or "skill" concerned, plus a label indicating the meaning of the two ends of the continuum, as was in fact used in the Foreign Service Institute Oral Interview in the 1970s for the performance factors Accent, Grammar, Vocabulary, Fluency and Comprehension:

1. ACCENT foreign __: __ __: __ __: __ native
(Wilds 1975)

The first attempt to provide detail about the kinds of behaviour associated with different parts of the continuum represented by the scale was the development of the "graphic rating scale" (Paterson 1922, Freyd 1923). The original graphic rating scales were a continuous line between two points. The dimension being measured would be described in a general definition at the top of the scale and behaviour associated with different parts of the continuum were described in short definitions called "cues." The cues would be spaced equidistantly along the continuum and connected to it so

as to present scale steps. However, raters were not asked to pick the most appropriate cue (or scale step) but rather, having decided which cue(s) were most appropriate to the performance being rated, to mark a point on the continuous line itself, which would necessarily be between cues.

The next significant development was to place the scale vertically rather than horizontally, thus allowing more room for longer and therefore more precise "cues" (Champney 1941). The difficulties of formulating precise cues and the danger of vague relative language which has been criticised in relation to scales of language proficiency (Brindley 1991, Alderson 1991a) was recognised in 1941:

> "Incisiveness: A cue must be more than mere words. It should describe behaviour with as much concrete vividness as is compatible with the breadth of the definition. The use of words like "rarely," "usually," "slightly," and "extremely" is only excusable if the scale value does not depend on them." (Champney 1941: 144)

Champney's second innovation was to promote a multidimensional model with a scale for each dimension and the recommendation that all subjects (talking of 5 or 6) should be rated on the same dimension before going onto to the next dimension. However, research since then suggests that this procedure still does not solve the problem of "halo effect."

The weak points of the graphic rating scale methodology, even incorporating Champney's innovations, were deciding the dimensions, selecting or designing the cues, and deciding what scale value to give the cues on each dimension. To assign values, Champney used a rank ordering rater-agreement task still often used in scale construction, for example in the development of the Eurocentres certificate scales. In this technique, the scale is cut up into its constituent descriptors, and these are presented to workshop participants who are asked to put them in rank order. Participants are then given the "correct" rankings and asked to highlight key concepts and formulations which helped or mislead them. Smith and Kendall developed a more rigorous quantitative methodology of cross-checking workshops and item analysis for identifying dimensions which made sense to the people who were to use the scale, for selecting the cues from a pool offered on the basis of consistent interpretation, and for giving scale values to those most appropriate cues. These behavioural cues "anchored" rating observations to the continuum—hence the name "Behaviourally Anchored Rating Scale" = BARS (Smith and Kendall 1963).

This is a different use of the term "anchor" to the use of that term in Rasch modelling, where anchors are items common to two tests, which link the tests into the same analysis.

Formats of BARS

BARS take various appearances, a classic simple one (Borman 1986: 103) gives a definition of the dimension concerned at the top, a scale on the left hand side, and a series of briefly worded examples of behaviour which could be expected at different levels of proficiency. No attempt is made to give a behavioural anchor to each step on the scale, nor do the "anchors" line up against scale points exactly. They are situated in the band between the points at the mean rating they received in the development workshops. The rater thinks of the behaviour which has been observed in relation to the "anchors," deciding whether this behaviour represents a higher or lower level on the continuum than each anchor, and selects the most appropriate scale step for the rating. A further Smith and Kendall innovation was to expand the qualitative description of the dimension which usually appeared at the top of the scale into three qualitative, more abstract descriptions: for a very high performance, for a very low performance and for an average performance, or minimum adequate competence. These three paragraph-long descriptors were then put on the left of the scale, at the top, in the middle and at the bottom. A variation on this original Smith and Kendall approach to qualitative descriptors is given by Landy and Farr (1983: 62–5). In this variant, the continuum is divided into three sections or broad ranges of level, with the qualitative summary statement covering the whole of each range rather than points at the two ends and middle; the behavioural anchors are also grouped in these three broad levels.

The BARS combination of, in our terminology, task information in the behavioural cues or anchors (*constructor/user-oriented*), plus qualitative information for broader levels (*assessor/diagnostic-oriented*), helped increase transparency. People could see what was being talked about, the potential of such a metalanguage for rater training, ratee feed-back, personnel training, job analysis, policy making etc. made BARS an increasingly popular evaluation format.

BARS Descriptor Style

A major consideration in the BARS approach is that the behavioural anchors refer to very concrete, specific examples of behaviour: the kind of thing a person at this level could be expected to do. Indeed the original name for the Smith and Kendall invention is "Behavioural Expectation Scales" (BES). The equivalence of such an approach for a scale of language proficiency would be a numerical scale with the anchors being tasks the student could be expected to perform on each of the dimensions at each of the levels. Following Smith and Kendall's style of wording, a language certificate statement like:

> "Can write on a range of subjects, and compose personal and straightforward formal letters so that, despite some errors and problems with formulation, the reader has little difficulty following." (Eurocentres Certificate, Writing, Level 6)

might form the basis of an "anchor" like:

> "If this student had to write to an English speaking friend in order to maintain contact and pass on some important information about arrangements, could be expected to produce a letter that the friend had little difficulty following, despite some errors and problems of formulation."

The appeal of BARS is the concreteness of the anchors. The main difficulty with the approach is that the raters have to judge where the behaviour observed "fits" on the scale. To do that, they have to *infer* how a person would behave in a specific given situation (on the basis of having observed them in *other* specific situations), and some people experience difficulty doing that. Notice that in the language example of the BARS approach fabricated above, focusing on one very specific behaviour to anchor a level of performance would mean that (a) the more general part of the certificate statement for this level "can write on a range of subjects" and (b) another specific behaviour mentioned: "can compose straightforward formal letters so that……" would both be omitted from the scale. This demonstrates the main problem people had with BARS: the examples tend to be too specific, and it can be difficult to generalise from them.

Successors to BARS

Although BARS continue to be used extensively, the response to these difficulties has been the development of two rival successors, which parallel very closely the directions taken in applying the ideas behind behavioural scales to language learning.

Behavioural Observation Scales (*BOS*): A long list of tasks in the domain, which are each separately rated on a numerical scale, usually 0–5. One variant is to just tick the tasks (Yes/No); another uses what is called a modified standard scale (MSS), giving a 0 for an average performance, a minus (-) for a low one, and a plus (+) for a high one (Saal and Landy 1977).

This is the kind of quantitative approach used in the graded objective movement. It is closely related to behavioural objectives in non-language vocational education (typing, machine skills), the so-called "competence-based approach" the "mastery learning" interpretation of criterion-referenced assessment. The trouble is that it does not place the student on a continuum in an overall framework; it is difficult to generalise about competence on the basis of checked-off assessments of performance on a list of tasks, particularly when, as is often the case, there is no comment about the quality of the performance (Ingram and Wylie 1989). Language cannot be effectively atomised in this way, and the specificity and partiality of the lists involved leads to the same kinds of problems with generalisability as were experienced with the original BARS, as mentioned above.

Such an approach could, however, be used within a defined broader level as a form of continuous assessment, and in most language applications this seems to be implicitly the case, since the applications tend to be in programmes aiming to get students up to The *Threshold Level*.

Behaviour Summary Scales (*BSS*): Anchor paragraphs are written which are representative of and common to the broader range of behaviours, incidents and subskills which are scaled at each level, and includes more abstract comment. In other words the kind of abstract comment about the quality of performance which started appearing on some forms of BARS is extended with examples of specific behaviours which are intended to be representative. The scale will probably have three or four sub-scales on different performance aspects, and it may group narrower numbered levels into broader defined levels (Landy and Farr 1983: 104–9). Like the original BARS, and unlike BOS, BSS scales thus take the definition of criterion-ref-

erenced assessment, not as mastering specific points in a domain, but as identifying someone's stage of development on a continuum (Hambleton 1988, Berk 1988, Glaser 1963).

One way this approach has developed in the language field is the user-orientated scales for different contexts of use pioneered by the LSP (language for specific purposes) scales of ELTDU (1976) and IBM France (1974 in Trim 1978), with a rating for each level for a set of specific language activities. This approach continues to be applied in the ALTE "Can do statements" under development. Another application is the scaling of different aspects of performance in assessment sub-scales for the degrees of skill required: *analytic* rating scales: (Carroll 1980; Shohamy 1981; Carroll and Hall 1985; Carroll and West 1989; IELTS; University of Cambridge/Royal Society of Arts 1990; Royal Society of Arts 1989; Eurocentres Assessment Scales).

The two approaches BSS and BOS differ in presentation, but all three types (BARS, BOS, BSS) tend to share the same development technique, usually simplified and less rigorous derivations of that employed by Smith and Kendall, and as a result to produce similar results. During the 1970s there were a whole series of inconclusive format comparison studies, some showing the one, some showing the other format to be superior. (See Jacobs et al 1980, Kingstrom and Bass 1981 for reviews.) Apart from methodological problems which make comparison difficult if not impossible, the inconclusiveness would appear to be due to the fact that the difference is mainly a presentational one, and superiority / inferiority is probably due more to rigour / lack of rigour during development (Borman 1986, Landy and Farr 1983). Borman and Landy & Farr therefore called for a moratorium on format comparison research and a concentration on deciding at what level to anchor the "anchors," which through inadequate item analysis often systematise the random error they are designed to exclude. The "anchors" should be anchored through a method based in psychometric theory and a lot of the problems experienced with behaviourally-based scales can be traced to the fact that they are not (Landy and Farr 1983). This is exactly the same criticism levelled against common standard setting practices in criterion-referenced assessment in language learning: as Clark said, they often seem to have been just "plucked out of the air" (Clark 1987: 44).

Advantages of Behavioural Definitions

Where the approaches are felt to differ is with regard to giving feedback. Behaviour Summary Scales, with the three or four sub-categories, provide a metalanguage for feedback, and a justification for decisions in a way that the seeming arbitrariness of the lists in BOS or the very specific selected anchors in BARS do not. Feedback is, indeed, felt to be the main pay-off from the behavioural anchors. Studies from the work performance evaluation field give, in general, only limited support to the idea that the transparency of defined criteria increase the *reliability* of ratings. A number of studies show the presence or absence of defined labels making no significant difference to means or reliabilities (e.g. Finn 1972, McKelvie 1978), especially when the content of the rating tasks is familiar and individuals have developed a common perspective and a set of similar "preconceived and rather uniform judgement standards" (Finn 1972: 264).

On the other hand, a number of other studies *do* show an improvement with definitions. Keaveny and McGann found that adding behavioural descriptors reduced halo effect and improved discriminant validity in a multi-trait multi-method analysis: raters found it easier to keep the dimensions being rated separate if they were given definitions. (Keaveny and McGann 1975). Borman and Dunnette conclude that scales with defined behavioural descriptors were clearly superior on inter-rater reliability, and classic rater errors like halo effect and leniency. However, they calculate that the addition of defined descriptors only increases reliability by at most 5% of the variance, and suggest that the true argument in favour of defined descriptors is rather the wealth of information which they furnish about overall requirements and about individual performance in relation to those requirements (Borman and Dunnette 1975).

A comprehensive review of these comparative studies came to the conclusion that the addition of behavioural descriptors "can no longer be considered the means by which rating errors are minimised." From a quantitative, psychometric point of view scales with descriptors are no better or worse than other methods, but the real potential is in the qualitative improvements that come from defining and giving feedback in relation to common goals (Jacobs et al 1980: 630). In a clarification of the original purpose of BARS, Bernadin and Smith declared that the aim had in any case been more *formative* than *summative,* to encourage accurate observation and recording of behaviour in continuous assessment in order "to enhance fu-

ture observation and to foster a *common frame of reference* in observers ratings" (Bernadin and Smith 1981: 458).

With regard to language learning, then, the main arguments for including detailed definitions of levels of different aspects of proficiency, either descriptions of tasks the learner can do, or aspects of competence, relate more to validity than to reliability.

Feedback. Detailed assessment categories supply a metalanguage for giving feedback on performance and suggesting areas to concentrate on. In a work performance context, it is estimated that the addition of transparent definitions of proficiency levels in feedback improves performance between 10% and 30% (Landy et al 1982: 21–23). Such a metalanguage can also be used in sensitisation activities with learners about the nature of language, and hence the nature of the activity of language learning. They can reveal that performance changes qualitatively with increasing proficiency, that assessment categories operate differently at different levels (Clifford 1980; Pollitt and Hutchinson 1987; Pollitt and Murray 1993).

Washback. Unfortunately virtually all support for washback effects seems to be anecdotal (Alderson 1991b). Nevertheless the principle of "washback validity" (Morrow 1986) or systemic validity (Fredericksen and Collins 1989) is very attractive. Experience in Eurocentres suggests that detailed assessment criteria sensitise teachers to the components of communicative language proficiency and for the necessity for learners to acquire skills of language *use* in group interaction. Since teachers are generally good critics, such discussion tends to develop into criticism of the activities which provided the assessment sample. Since the language generated by material depends on atmosphere and how the activity is set up, this invites discussion of the management of a communicative classroom. This in turn invites discussion of classroom discourse and the phenomenon of TTT (Teacher talking time): is the classroom a place where questions are asked, answers displayed as knowledge, and responses patronizingly rewarded (Sinclair and Coulthard 1975; Sinclair and Brazil 1982) or are there opportunities for autonomous interaction?

Uniformity of Scales and Grids

Most scales and grids have a verbal descriptor for all levels or bands on the scale. If the scale has a number of situational categories (as with IBM

France) or qualitative categories (an analytic scale like Shohamy 1981), it makes up a grid in which all the boxes (each category for each level) are filled. As mentioned in Chapter 1, Pollitt & Murray (1993) have argued that, in relation to qualitative categories, such full grids are diagnostic-oriented and that a more differentiated approach bringing out the assessment features salient at different levels (as a holistic scale does) might be more appropriate for assessment.

There are in fact a number of ways in which the uniformity of a filled grid can be varied. This is not necessarily to say that a full grid is a "bad thing" but to point out that in the same way that a "scale" can in fact be a multitude of sub-scales (i.e. multidimensional in that sense), the descriptors used for different levels on a scale can be non-uniform.

Not all scales which have defined points or bands would require one to place a learner (or examination) exactly on a band. The very earliest behavioural rating scales had descriptors or "cues" for each scale point (Champney 1941), but individuals were rated anywhere on a continuous line, i.e. *between* the scale points. It is a commonly observed phenomenon that language teachers often want to do this with rating scales to reward achievement or to rank their class (norm-referencing). Comparisons of examination levels may be more informative if they are placed relative to another against a common yardstick at the points where they appear to fall rather than being placed neatly in a band suggesting exact equivalence (North 1992a: 163–5).

Descriptors between Scale Points

Scales need not have descriptors placed on the points of the scale; i.e. scales do not need to have "bands." Early forms of Behavioural Expectation Scales (BES) departed from Champney's practice by placing the "cues" where they landed empirically, i.e. between bands (as suggested with the examinations above). These "cues" were examples of expected behaviours. This (task) information was often supplemented by holistic, qualitative statements defining the top, mid-point and bottom of the scale, or broad ranges of levels (top third, middle third, bottom third). Candidates were then rated *onto* the scale points. Bachman and Palmer's (1982) scale for language competence places some descriptors on scale points, others between them, and others covering a band of levels.

Descriptors Missing

Scales need not have descriptors for all categories for all levels. The ELTDU scale contrasts with the IBM France scale in that entries are made only when they are appropriate. Thus there are blanks at the top, and the bottom or even in the middle of the sub-scales. Definitions are not forced. A blank in the middle means that the previous definition still applies at this band because a performance cannot be distinguished from the performance for the band below. The CASE (Cambridge Assessment of Spoken English) scales (assessor-oriented) started with entries for all qualitative categories for all levels but through empirical analysis of the way the scale was used, reduced the scale from 9 to 5 bands for most categories, with only 4 bands for "interlocutor support" and "task achievement."

Changes in Categories

Scales need not use the same categories at all levels or broad ranges of level. An assessor-oriented scale could use different qualitative criteria at different level ranges, focusing on aspects found to be salient to raters at those levels (Pollitt and Murray 1993). A common framework scale could employ different qualitative criteria for different educational sectors (North 1992b). The National Language Standards developed by the UK Languages Lead Body (1993) has a holistic qualitative descriptor for each of the four skills for each level, which is titled "commentary" but each skill is then subdivided into a number of functional categories: language activities called "elements." These categories are not the same for each level, though there is some overlap. Each element then has its own appropriate constructor-oriented and assessor-oriented information, set out to a uniform pattern.

Variable Scale Steps

Scales need not always use the same size of level unit. Trim (1978: 51–54), Clark (1985: 142–3) and Ingram & Wylie (1989) discuss ways in which grades can be attached to broader framework levels. As suggested when talking about the mapping metaphor, there may be occasions when a switch to smaller units, more akin to learning modules, may be applicable. In the GLAFLL project, such smaller units were called "waystages" and according to Clark, each school developed their own more or less independently; standardisation was confined to assessments at the framework levels (called "stages"). Different sets of waystages or mini-scales using appropriately

sized steps could be developed for different educational sectors in relation to a common framework scale (North 1992b), and local or institutional mini-scales or waystages could be calibrated to a common framework scale used mainly to report harder (examination) data (Hargreaves 1992).

The Languages Lead Body (1993) National Language Standards are a good example of the way a meta-state (profiling) criterion-referencing model and a continuum (scalar) model can be reconciled around a notional global scale. Although the system is clearly conceived as being a modular unit/credit scheme, various aspects define the levels themselves in scalar terms. Apart from the published Levels Descriptors which provide a global scale of proficiency, and apart from the Commentary attached to each Unit (each of the four skills for each level) which make up a kind of holistic scale for the skill in question, there were at least two other scalar aspects of the system apparent during development. Firstly, "Draft Attainment Standards" (*constructor/user-oriented*) in the form of a list of "Can do statements" in 9 key areas for each level were drawn up by Chris Yates (ELTDU) to be used (e.g. by training managers) to specify the content of specific language requirements. It is not difficult to imagine these statements presented as 9 sub-scales, with blanks for the top and bottom level on some of them, precisely as is the case with the ELTDU scale. Secondly, in the July 1992 in the last internal edition, there is an analytic grid defining the quality of performance expected at for 5 "factors" (*assessor/diagnosis-oriented*). This grid was used to help draw up performance criteria and range statements (to do with transferability) when formulating the text for each standard. Although not included in the final document, these two *insider* tools demonstrate the scalar thinking behind the modular system.

Dimensionality

The use of the word "scale" sometimes causes people to overlook the fact that what is referred to in both language assessment and in the evaluation of work performance (BARS, BSS etc.) is often an instrument covering several dimensions, a family of sub-scales covering different categories, and therefore *in this sense* multidimensional, and that if a global (aggregate) scale is offered, this is only a summary of a more complex picture. The "scale" is a ruler ("yardstick") used to give a vertical axis, and hence create a matrix to take account of proficiency in different domains, circumstances, activities, skills, etc. A profile drawn across the dimensions of a scale matrix might be presented a telephone number e.g. 4544324 4534624 45553444 (North et al

1992), or if three or more axis are envisaged (e.g. context of use, qualities of performance, degree of skill) as something more like a correlation matrix, or as a circle or series of circles. Whichever is the case, such presentations giving a profile assume a common vertical axis or dimension, and imply common units of measurement on the different sub-scales even if they are used at different degrees of delicacy for different domains as on a real geographical map.

Creating in a principled manner a profile grid which uses the same sized steps for different categories is certainly a complicated process. Writers who criticise the use of scales of language proficiency often start from the assumption that language is clearly multidimensional. This is to assume that producing a scale or grid for language proficiency with common steps would violate this multidimensionality—as if language were in some way special, different to other subjects. In fact it is the case that virtually everything that is measured is multidimensional, and many subjects are far more multidimensional than language ability. For example, according to Linacre (personal communication) in mathematics, higher level skills are of a quite different nature to primary school numbers and arithmetic, and at secondary school ability in algebra, geometry, arithmetic and logarithms can be relatively unrelated since they call on quite different competences and mental processes.

The distinction between multidimensionality and unidimensionality is not nearly as clear-cut as many people imagine. "Unidimensionality (the notion of a single continuum) is a relative concept and is constructed either to understand complex phenomena or to facilitate decision-making" (Andrich 1988: 303); it is "conceptual rather than factual, qualitative rather than quantitative" (Wright and Linacre 1989: 3), "a matter of degree" Choi and Bachman 1992: 74). It is measurement or judgement that is unidimensional, not the thing being measured, and this fact remains whether one is talking about the kind of "okay/not okay" distinction made by examinations or in the rather mechanistic interpretation of criterion referencing in "mastery learning" or in plotting attainment onto a continuum of stages. It is no contradiction for the subject being studied to be *psychologically* complex and multidimensional whilst still exhibiting *psychometric* unidimensionality (Henning 1992) and there is a wealth of evidence that suggests that, in psychometric terms, language ability can be regarded as unidimensional and every test or assessment which reports one final result assumes that it is (Henning 1988: 83).

Conversely, findings of psychometric undimensionality, e.g. through correlations and correlation based techniques like factor analysis are no evidence for a single unidimensional language ability (Unitary Competence Hypothesis: Oller 1976/83). In fact data from children's weight, mathematics achievement and reading ability at various ages between 6 and 16 would in fact be highly intercorrelated "suggesting" psychometric unidimensionality (Upshur and Homberg 1983: 194). Psychometric unidimensionality is a complex phenomenon; data requires careful investigation to identify the correct interpretation. After applying four tests of unidimensionality to a placement test and getting conflicting results Blais and Laurier conclude:

> "These findings suggest that unidimensionality is not itself a unidimensional concept at all! The concept is defined through approach. The issue that is now raised is not simply which approach is the best one but what kind of unidimensionality we look for. …Unidimensionality is not a yes/no issue; it is rather a matter of degree considering the *purpose of the test*." (Blais and Laurier 1993: 13–14)

Stahl comments:

> "It is the *intention* of the researcher which defines the dimension…. One researcher's multidimensional space of incongruities may be another researcher's unidimensional line of concern." (Stahl 1991: 265)

Seeing the issue from an applied linguistics rather than a psychometric viewpoint Davies adds:

> "…The problem is…*a*…*philosophical* one of whether a distinction between a unitary and a non unitary competence has any meaning. It appears that is possible to demonstrate from the data we have that either conclusion is correct depending on the type of analysis of the data we use. In other words both the UCH and the no-UCH are "correct" since they reflect different ways of approaching the same issue. They are both right as we can see on the grounds of common sense in that, at some level, there is a unitary language skill, the level at which the distinctions among the performance skills of speaking, writing etc. are unimportant. But at some other level, these very distinctions become very important when we consider issues such as illiteracy and being better at say speaking than at reading. The issue therefore of whether one or the other is correct becomes a non-

> theoretical issue, while of course remaining very much a practical one." (Davies 1991)

Finally, in a class room teachers' guide to testing, Baker puts the matter more pragmatically:

> "…there are times when we may be interested in assessing language proficiency in a general way and not worrying too much about its structure or the content of the test. The placement of learners in a general language instruction programme is such an application." (Baker 1989)

To paraphrase Davies and Baker: the fact that people's individual abilities vary substantially across skills and/or across aspects of competence does not alter the fact that it sometimes makes sense to just consider a global summary outcome. A good example of when this is done is when deciding which class to put a student into in an intensive language programme. The classes are organised along one dimension: ability, the ordering is thus unidimensional.

One could add that even here in a placement context, the quality of the information available to the class teacher and student would be improved if one could give feedback as a profile, to assist matching current position to target profile. But the point being made by Baker is that in most placement contexts the prime decision, which class, is based on one dimension: ability. The profile, if there is one, is added detail. Such a profile (for entry or exit testing) will only make sense if it can be mentally matched to that same dimension. A profile (whether of a person or of a programme) can only be seen as a profile if it is contrasted against something; that "something" should logically be a "non-profile" i.e. something standard, common, in order to highlight what is non-standard, particular. Otherwise the profile will cluttered with unnecessary complexity which makes it less usable. This non-profile is the single dimension you get by adding up and averaging out all the profiles; it is the flat, abstract profile, the "global scale" in the sense in which Carroll (1983: 83 and 94) explained global proficiency.

Therefore, the wording of dimensions on a profile should not be developed by translating/word-processing descriptions of "some neat and tidy intuitive ideal" (Clark 1987: 46) across skills (active to passive) or across contexts (general to specific as in ESU: Carroll and West 1989), but should rather be developed in relation to realistic norms of classroom achievement for the group(s) in question (Stern 1989: 214). As mentioned in the discus-

sion of norms and criteria above, this posits an empirical basis to the development, which can be provided by the Rasch Rating Scale Model.

If it is suspected that learners' abilities in certain content areas may develop in a substantially different manner, then steps can be taken to discover this during the course of the analysis, for example by looking at fit statistics and standardised residuals (Hambleton et al 1991: 66) and by examining the stability of estimated values when the area is analysed separated compared to when it is analysed in the full data set (Bejar 1980). In certain cases it may turn out to be necessary to remove areas from the analysis, as proved in fact to be the case with this study. If it is suspected that different sub-groups of learners may react differently to one another in relation to certain groups of items, then again this can be investigated empirically. The view taken in this study is that these issues of dimensionality can be and should be investigated empirically in relation to the classroom achievement of the kinds of learners concerned.

Types of Measurement Scales

Classical measurement recognises four different sorts of scales, in order of rigour, which might, in theory, be used for this purpose (Keeves 1988: 404):

- Nominal scales which categorise data (e.g. nationality);
- Ordinal scales which organise data in rank order;
- Equal interval scales which use a mathematically constant unit of measurement to give steps the same size;
- Ratio scales which have an absolute zero as a starting point, and a mathematically constant unit of measurement.

A type not listed by Keeves is a scale which is linear, on which the size of the steps are known and arithmetically in proportion, but on which the intervals are not equal. This would be in the third place in the list between ordinal and equal interval scales.

Nominal Scales

The "weakest" scale is nominal scale: a modular profiling system in which the modules do not have a hierarchical relationship to each other and can, in theory, be taken in a different order. This is the principle behind the DELF (Diplôme d'Etudes en Langue Française), although it should be

noted that the choice of order is taken at a *national* level when the syllabus is developed; there is no individual choice.

Ordinal Scales

Most existing scales of language proficiency can probably be regarded as ordinal scales. However, it should be noted that for this to be strictly true, it should be possible to demonstrate that people rank the descriptors in the same order with a high degree of consistency. That people will not be able to do this with 100% consistency is to be expected: a "flip-flop problem" (Bernadin and Smith 1981: 461) in which people get descriptors reversed, is impossible to eradicate completely since the more people are asked to do the task, the higher the probability of disagreement. In any case, most scales of proficiency are certainly not more rigorous scales than ordinal scales, so when using classical test theory, statistics suitable to ordinal data should be used in relation with them (Bachman 1990a: 45).

Linear Scales

This is a type not usually mentioned in traditional classifications, like that of Keeves cited above. Raw data from proficiency scales, like that from tests, is often treated *as if* it was linear both in statistical analysis and in decision-making, when in fact it is not. The result is error. One way to avoid this error is to convert from raw to standard scores like z-scores which will report on a linear scale centred on zero. The logit scale produced by a Rasch model or other IRT (Item Response Theory) analysis is also centred on zero and is also linear whilst not being strictly speaking equal interval "although it is popular and reasonable to assume that (it) has equal interval properties" (Hambleton et al 1991: 87).

Rasch offers the only currently available way to provide an objective linear scale (Wright and Linacre 1987: 2); and hence "scale-free measurement:" measurement that transcends the brand-name of the instrument used (Fisher 1993: 273). A linear scale of language proficiency can be produced with a Rasch model analysis of test scores (e.g. Adams et al 1987; Brown et al 1992; Jones 1993) or ratings (Griffin 1989, 1990a, 1990b) as in this study. The real size of bands on an existing scale of proficiency for different categories can be determined with the Rasch Rating Scale Model (Wright and Masters 1982) e.g. Milanovic et al (1992/6).

Equal Interval Scales

Often just called interval scales (e.g. Bachman 1990a: 28) equal interval scales are linear scales on which the steps are of equal size, like on a ruler or thermometer. It is highly unlikely that any existing scales of language proficiency meet this criterion. The same reservation applies to all test scores. The fact that scales of proficiency do not have equal intervals does not matter provided everybody knows this. After all it is widely accepted that language learning, in common with other kinds of verbal learning, does not progress in a constant linear fashion—but rather by progression through a series of plateaux (Bryan and Harter 1899 cited in McDonough 1981: 35).

The problem is that people tend to treat scales of language proficiency *as if* the steps were equal in relation to seat time, even though the authors often go to some pains to try and point out that this is not meant to be the case by emphasising that the steps increase in size as one goes up the continuum (e.g. Carroll B.J. 1980: 86–7; Carroll and Hall 1985: 77; Higgs 1984: 6; Liskin-Gasparro 1984a: 477; Liskin-Gasparro 1984b: 26; Lowe 1985: 21–3; Clark and Lett 1988: 78). This was a criticism of the British Council ELTS scale (Carroll B.J. 1978/1983) that subjective ratings were "used as if they are genuine experimental figures on a true equal interval scale about which conclusions can be drawn on the length of study needed" (Criper 1981: 118). If it is necessary to provide information about how long a certain type of student in a certain type of setting might take to reach a certain objective from a certain starting point, then such information should be based upon experience and a minimum of empirical evidence—not whatever may have intended about units of seat time at the design stage.

Ratio Scales

Ratio scales have a origin of absolute zero. It is worth pointing out that "many measurement applications in the physical science, such as pitch, hue, loudness, hardness, do well without any absolute zeros. Neither length nor time have natural origins or absolute zeros" (Bezruczko 1990: 115). When measuring a table, the floor is a convenient starting point, but when measuring a mountain sea level is more convenient (Linacre 1989: 64). Measurement requires counting standard (hence abstract) units from an agreed (standard) starting point. Whether the zero is absolute or not does not matter provided it is an accepted standard (Wright and Linacre 1989: 2).

Hence, as Bachman points out, the Fahrenheit and Celsius scales are not ratio scales since zero is placed at a point of convention.

In measuring language proficiency, we are hardly aiming at a greater precision than that required by physical sciences. Yet Bachman comes to the conclusion that points of absolute zero and perfection are prerequisites for criterion-referenced testing (Bachman 1989b: 254–6), and that a common framework scale should have such an absolute zero and perfection and be defined abstractly in terms of the presence or absence of certain features (Bachman 1990a: 344–6) This is a very strict interpretation of Glaser's original concept of criterion-referenced assessment which would disqualify virtually all existing applications of it. It is equivalent to saying that criterion-referencing in general and a common framework scale in particular must have a ratio scale although existing common framework scales dealing with the physical world such as temperature scales, cartography conventions and common scales for storms, earthquakes etc. etc. do not.

The kind of scale appropriate for a common framework thus does not fall into any of the four classical categories, but is rather an objectively calibrated linear scale, No 3. above, such as can be provided by calibrating descriptors with an Item Response Theory measurement model like the Rasch model.

Essentials of a Valid Measurement Scale

To be effective, however, any measurement scale should satisfy the following points listed by Thorndike (1904/1912: 5 cited in Engelhard 1991a) as essential for a valid scale:

1. Objectivity
2. Consistency
3. Definiteness of facts
4. Comparability with the facts to be measured
5. Reference to a defined zero point

Objectivity

Objectivity "is the requirement that the measures produced by a measurement model be sample free for the agents (*here: scale*) and test-free for the objects (*here: people*)" (Wright and Linacre 1987: 2). Thurstone has defined two ways in which a scale should be "objective." Firstly:

> "The scale must transcend the group measured…its function must be independent of the object of measurement." (Thurstone 1928a, cited in Wright. 1988: 3)

This is another way of expressing the point made by Wright and Linacre that the scale should be sample-free. Secondly, the values attached to the scale:

> "…must be as free as possible, and preferably entirely free from the actual opinions of individuals or groups." (Thurstone 1928b, cited in Wright and Masters 1982: 5)

In particular:

> "…the scale values of the statements should not be affected by the opinions of the people who helped to construct it. This may turn out to be a severe test in practice, but the scaling method must stand such a test before it can be accepted as being more than a description of the people who construct the scale. At any rate, to the extent that the present method of scale construction is affected by the opinions of the readers who help sort out the original statements into a scale, to that extent the validity of the scale may be challenged. " (Thurstone 1928b: 547–8, quoted in Wright and Masters 1982: 15)

It is a disadvantage of existing scales of language proficiency that they cannot meet these requirements. As was pointed out in the discussion of the structure of language proficiency all classical statistics are sample-dependent. Therefore scales which can demonstrate a modicum of empirical validation using classical statistics (e.g. ASLPR: Ingram 1985; Eurocentres: North 1991; ACTFL Thompson 1995) cannot strictly claim any validity beyond those particular administrations which were studied (Engelhard and Osberg 1983: 283; Henning 1984: 124; Henning 1987a: 108–9; Lord and Stocking 1988: 269). These requirements can be satisfied with the Rating Scale Model version of the Rasch model from the Item Response Theory series of probabilistic models (Wright and Masters 1982; Masters 1982; 1988a; 1988b; 1990).

In this context, it is not surprising that there is an increasing tendency to employ the Rating Scale Model, and particularly the many-faceted version of it (Linacre 1989) in validation studies of existing scales of language proficiency (e.g. ACTFL: Stansfield and Kenyon 1992; ASLPR: Lee et al 1998).

Consistency

Consistency is glossed by Engelhard as reliability in the classical sense, an ability to separate subjects on the ability continuum, a concept which is related to psychometric unidimensionality. "The series of facts used as a scale must be varying amounts of the same sort of thing or quality" Thorndike 1916: 13, cited in Engelhard 1991a). As Swain (1993) points out in relation to the DBP project, psychometric theory requires a test to be relatively homogeneous. To the extent that it is not, low reliability estimates will be produced by classical test theory procedures. The argument is that if the test is not relatively homogenous, it should be subdivided into sub-tests until those sub-tests are relatively homogenous. Conversely, sub-tests which inter-correlate highly should be added together as they are producing psychometric homogeneity. Lots of sub-tests which are not homogenous, and which inter-correlate as in the DBP project make little sense.

The same argument can be applied to scales. Sub-scales only make sense if the descriptors are consistent: i.e. if they properly belong on the same sub-scale. Otherwise the sub-scale is not measuring a defined dimension. Such unidimensionality of sub-scales can be checked by screening the pool of descriptor items to ensure that those grouped under the same dimension are perceived to be on that dimension by raters. The Smith and Kendall (1963) technique to do this has been discussed. Most computer programs for the Rasch model provide separability statistics which are similar in function to a KR 20 estimate of reliability in the sense of separability (Wright and Masters 1982). In addition there are methods based on comparing values for different subsets of items (after Bejar 1980) and on examining the difference between what the model predicted and what actually happened, using statistics provided for each item (residuals, fit statistics). This is discussed in more detail later in the chapter.

Definiteness of facts

This is Champney's "incisiveness." This is the reason for the very concrete tasks used by Smith and Kendall (1963) as anchors in BARS, the reason for the length and detail of the ASLPR descriptors, which aim to help the rater fix the "type" in mind (Wylie 1993, personal communication). Ingram states that it is to help with this problem that the ASLPR adds, separately, lists of example tasks the learner can perform in addition to the main, more abstract description of competence for each level. He suggests that this may

be why the ASLPR reports higher reliability that other scales (Ingram 1985: 254). As mentioned under point 1, however, allocating concrete tasks to particular levels is problematic if it is done by copying conventions in other scales, rather than on the basis of consensus supported by empirical evidence. The effect is to systematise error and produce invalid measurement. Smith and Kendall's technique to avoid the problem has been mentioned. The disadvantage of their approach was that it tended to reject descriptors in the middle range (where there is always more disagreement, and where classical statistics work less well to separate items), but it avoided systematising error in the scale. Subsequent behavioural scales (of all formats) were often less rigorous in their approach, offering therefore a spurious definitiveness and thus less reliable ratings (Murphy and Constans 1987; Murphy and Pardaffy 1989), which made meaningful comparison between formats impossible (Kingstrom and Bass 1981; Landy and Farr 1983; Borman 1986).

Scaling aspects of proficiency presents the problem that one may either fall into the trap of naively assigning forms to particular levels, or as Alderson (1991a: 81–2) admits, one may be forced back on juggling qualifiers, the opposite of Champney and Thorndike are saying. As one applied linguist has put it, one gets the impression that some scale authors have decided on categories and then spread them evenly out along the continuum with qualifiers. The "absolute" scales recommended by Bachman (Bachman and Palmer 1982; Bachman 1990a: 326–8) are extreme examples of this tendency (Brindley 1991: 11) as are those for the Cambridge Assessment in Spoken English (CASE). Such scales rely upon raters remembering from common rater training what is supposed to be meant by the labels. Such an approach fails to meet Thorndike's requirement, and while it may be practical for an examination which can get its raters together for such training (like CASE), it is difficult to see how it could be feasible for a common framework. For the scale to be interpretable by a wider group of users, statements of more definite behaviours are certainly desirable.

Table 3.1 compares the two sub-scales for *Vocabulary* in Bachman and Palmer (1982) and in the CASE (1992) scale. Much hinges, in both scales, on the difference between the words "extensive" and "large" or between "small" and "limited". The optical impression is that the CASE "D" (Restricted range: basic communication) is intended to mean much the same as Bachman and Palmer's "Small Vocabulary" reinforcing the question as to what the difference is between a "small" and a "limited" vocabulary.

Table 3.1: Minimalistic Definition Styles

Bachman and Palmer		**CASE**	
4	Extensive Vocabulary	A	Extensive range; accurate and appropriate; wide topic range.
3	Large Vocabulary	B	Large range of everyday vocabulary accurately used.
2	Vocabulary of Moderate Size	C	Moderate range for everyday use.
1	Small Vocabulary	D	Restricted range sufficient for basic communication only; choice limited.
0	Limited Vocabulary (A few words and formulaic phrases)	E	Very restricted range usually inadequate for clear communication.

Comparability

Comparability with the facts to be measured (i.e. appropriacy in context). This means firstly that the scale should be appropriate to the domain and context in which it is being applied (Spolsky 1986: 150; 1993: 208; Spolsky 1986: 154; 1989: 65; Brindley 1991: 154–5). Secondly the orientation of the scale must fit the purpose to which it is to be put (Alderson 1991a). A scale which says what kind of tasks the learner can perform in general is *constructor-oriented* (to help design a syllabus or assessment activity), and/or *user-oriented* (to report this range of capabilities). A scale which defines aspects of proficiency used to evaluate learner performances on a range of tasks is *assessor-oriented*. A master grid from which such assessor-scales could be extracted (taking categories most suitable to the task(s) in hand) would be *diagnosis-oriented* (Pollitt and Murray 1993). The Eurocentres framework keeps the types of information separate; the ASLPR puts them in different columns; the ACTFL/ILR system mixes them up and displays a poverty of assessor- or *diagnosis-oriented* information which has been criticised by Pollitt:

> "If as in the ACTFL writing scale, we find descriptions of stimulus (task) when we expect descriptions of response (performance), I at least feel a distinct lack of definition: where are the criteria for acceptability, and why are they being kept secret?" Pollitt (1991: 88)

A Defined Zero

There has to be a standard place to start counting but as for temperature, it is irrelevant whether it is an absolute zero or not provided, as Thorndike insisted, it is publicly known whether the starting point is defined. This requirement can be met by applying a Rasch model during the development of a scale since "zero" will then be defined by default at the exact centre of the scale, whilst "perfection," infinity, is to be found at the two ends. As Bachman points out (1990: 345), one can, if one wishes, define a certain ability level as zero. This could be a "mastery learning" cut-off point on the difficulty continuum, but one could also define a particular learner or mean for a group of learners (held to be prototypical for a particular educational threshold) as zero on the ability continuum. In relation to the many-faceted Rasch model when one is interested in judge-severity, one can define the rating of an expert judge or mean average of a panel of expert judges as zero on the judgement continuum, in order to highlight variation from it (Engelhard 1992: 15).

Common Methods of Scale Construction

A number of scaling methods which can be used to create a scale of language proficiency have been developed. Most involve small workshops with groups of informants.

Piles by Level

Informants sort work samples (or descriptors) into piles representing the number of levels they can distinguish (Thurstone 1928b).

Piles by Category and by Level

Informants sort descriptors into piles according to categories they are supposed to represent, and according to their difficulty. Categories are fine-tuned, and descriptors accepted/rejected on the basis of the number of misplacements and the size of standard deviations and scaled at their average difficulty (Smith and Kendall 1963).

Comparative Judgements

Groups rate pairs of performances stating which is better and why, in order to identify the salient features in performances at each level, that can then

be formulated into descriptors (Pollitt and Murray 1993, citing Thurstone 1959, Kelly 1955).

Key Concepts

Experienced raters identify key sample scripts for each level, and then agree the key features of each script. Features felt to be characteristic of different levels are then identified in discussion and incorporated in the descriptors (Alderson 1991).

Primary Trait

Informants sort performance samples into rank order; a common rank order is then negotiated, the principle on which the scripts have actually been sorted is then identified and described at each level, taking care to highlight features salient at a particular level (Mullis 1980).

Discourse Analysis and Multiple Regression

Performances are analysed in terms of occurrences of certain features whose contribution to the quality of performance is analysed with multiple regression. Features are worded into scale bands descriptors which are then validated (Fulcher 1993).

Each of these methods can be useful in the development of a scale of proficiency for a particular context, and indeed the first three were employed during the workshops with teachers in this study which are described in Chapter 4. However, in terms of the development of a common framework, these methods share two drawbacks in relation to the requirement for objectivity set out earlier in the chapter.

The results are context specific, determined by the particular characteristics of the performance samples used and the opinions of the informants used to make judgements. This means that the scale will not "transcend the group measured" (Thurstone 1928a cited in Wright 1988: 3).

The scale values of the statements are, with the exception of Fulcher's method "affected by the opinions of the people who helped to construct it" (Thurstone 1928b: 547–8, quoted in Wright and Masters 1982: 15).

The Rasch Measurement Model

The Rasch Model (Rasch 1960/1980 cited in e.g. Wright and Masters 1982; Masters 1990) is a simple one parameter model of Item Response Theory

(IRT) a branch of Latent Trait Theory. Rasch has been described as "the most important advance in psychometrics since Thurstone's 1927 law of comparative judgement" since "items are calibrated and persons are measured on a common interval scale" (Wright 1988: 286) thus providing an opportunity to mark out "a continuum of increasing proficiency" (Masters 1990: 58) so that "all possible scores on all possible tests are automatically equated in the measures they imply through the common calibrations of their bank items" (Wright 1988: 287; 289). Rasch thus offers a way to meet the requirement for objectivity identified above, since "objectivity is the requirement that the measures produced by a measurement model be sample free for the agents (test-items *here descriptors*) and test free for the objects (people)" (Wright and Linacre 1987: 2).

Characteristics

The model is called a one parameter model because it only deals with one parameter: difficulty. There is also a two parameter models taking account of the discrimination of items in addition to difficulty, and a three parameter model also taking account of guessing, in for example multiple choice tests. Because the two and three parameter models are extremely complex, not very robust and difficult to work with, the vast majority of IRT applications use the Rasch model. Good, short overviews of the way Rasch can be applied to language teaching are available from Henning (1984), Woods and Baker (1985), Pollitt and Hutchinson (1987) and Baker (1997).

The Rasch model uses a linear scale of the logarithm of the probability that a person will succeed on an item, or that an item will be answered correctly. This value is computed on the basis of all the decisions made in the rating or testing in "logits." Because the logit scale is a logarithmic scale, zero falls in the middle of the scale, which progresses towards infinity on either side. Hambleton, Swaminathan & Rogers (1991) state the IRT ability scale "may be thought of as an absolute scale with respect to the trait or ability that is being measured" (1991: 79) and suggest the term *proficiency level* as an appropriate term to describe someone's position on the scale. They point out that, contrary to popular belief, the IRT ability scale is *not* a ratio or equal interval scale "although it is popular and reasonable to assume that the theta scale has equal-interval scale properties" (1991: 87).

Reservations and Technical Complications

Certain reservations were expressed about the educational implications of the Rasch model when personal computers made it first available. These reservations relate to two main points. Firstly, it is sometimes assumed that the calibrations of an item bank, once established, are true for all time. In fact curriculum developments over a period of years may mean that some items get "easier" and others "harder" (Goldstein 1981, Tall 1981). This is in fact not a problem confined to the Rasch model, but a problem of any standardised scoring system. The Rasch model actually facilitates recalibration checks to detect value shifts over time. Secondly there is a danger of assuming that the calibrations of an item bank developed with one population (e.g. Chinese schoolchildren) can be applied to another (e.g. Scandinavian adults) (Goldstein 1981, Tall 1981). This is a delicate problem since it requires a value judgement about the point at which a group ceases to be a variant of the same population, and becomes a new population. However, the answer can be established by doing an independent analysis of the group in question, and then adding the new group to the main data set to see if there is a significant problem by comparing results (Henning 1988: 98; Linacre 1992, personal communication). Differential item functioning in relation to particular groups can also be quantified.

A third reservation relates to the choice of IRT model: Rasch is the simplest and most widespread model, but there are the two others. Both Choi and Bachman (1992: 74) and Hambleton et al (1991, Chapter 4) recommend trying the model fit to data of all three types of IRT models. Whilst this may be desirable, the prerequisites of using the two and three-parameter models (e.g. 1,000 subjects per test/questionnaire) were not feasible in view of the scale of this study, in common with many language testing projects (Madsen 1986: 2).

In any case, as Wright says (1992: 200) Rasch is a theory-based measurement model, not a descriptive model, and therefore it is necessary to demonstrate that a particular data set fits the assumptions of the measurement model, not that the measurement model fits the data. The measurement model is not there to describe the data. It simply should not be used with data which does not fit its assumptions. What else may be done with that data is another issue; Wright would argue that whatever it was would be a waste of time because if the data does not fit the rigorous requirement of the Rasch model—it is not very good data. He criticises other IRT models (2 and 3 parameter) because they tend to assume that the

data makes sense, and are "patched up to chase after whatever comes up" (ibid: 200)—very often from responses to multiple-choice questions. He therefore does not accept that the 2 and 3 parameter models provide objective scaling since results are distorted to make bad data fit (Personal communication).

All Choi and Bachman's arguments are, in fact, based on data from multiple choice items, which are a somewhat discredited type of test item since they test (possible) recognition rather than use and since students develop "test wiseness:" guessing strategies (Allan 1992). Since "no guessing" is one of the three stated assumptions of the Rasch model (see below), it should therefore come as no surprise if multiple choice data does not appear to fit Rasch very well.

Assumptions

The Rasch model fits items and persons onto a linear scale by making three assumptions: unidimensionality; local independence; no guessing. In fact the model is relatively "robust" in that it can cope with a certain degree of violation of all three assumptions (Forsyth et al 1981).

Unidimensionality has been discussed already in relation to global language proficiency scales at the end of Chapter 2. It is also a requirement for the successful application of an IRT model. As has already been mentioned it is essential to distinguish between *psychological* unidimensionality and *psychometric* unidimensionality (Henning 1992), which is a relative construct (Andrich 1988: 303; Wright and Linacre 1989: 3; Choi and Bachman 1992: 74) dependent on the intention of the test and/or researcher (Stahl 1991: 265; Blais and Laurier 1993: 13–14). For a Rasch analysis, sufficient *psychometric* unidimensionality is required. Henning et al (1985) demonstrated that the common language skills division: Listening, Reading, Writing (Error-recognition), Grammar, and Vocabulary can all be accommodated within the Rasch model. Blais and Laurier (1993) demonstrate that a test will appear uni- or multidimensional depending on the "test" of unidimensionality used. In terms of profiling, Pollitt and Hutchinson's (1987) study shows that tasks which are socio-linguistically sufficiently distinct to produce radically different results can also be accommodated by Rasch. The Rasch model requires "a *dominant* component or factor that influences test performance" (Hambleton et al 1991: 9) and as Bachman has pointed out in summarising the state of play on the UCH (unitary competence hypothesis) proficiency is

currently considered "of a number of specific abilities as well as a general ability or set of general strategies or procedures" (Bachman 1991: 673).

The important point about unidimensionality in relation to the Rasch model is that it can be checked. Unlike conventional test statistics, which rely equally on unidimensionality (Henning 1988: 83), the Rasch model *tells you* when sufficient psychometric unidimensionality is not met in the data with fit statistics and residuals. After dismissing most conventional methods of testing for unidimensionality (which tend to produce different results as Blais and Laurier demonstrated), Hambleton et al (1991: 56–8) recommend a range of techniques for testing model-fit, among them (a) the so-called Bejar method (1980), which has been criticised and defended in lively correspondence in Language Testing (Spurling 1987a; 1987b; Henning 1987b; Henning 1988; Henning 1992), (b) comparing the stability of ability estimates using easy and difficult items (not used in this study), (c) checking the stability of item difficulties using different population sub-groups, and, (d) studying the relationship of residuals and standardised residuals of model fit to data fit, which they consider to be the best approach (1991: 66).

Local Independence and Guessing. The other two Rasch assumptions are not very relevant to the current discussion. The Rasch requirement for local independence means that items should not be dependent on one another: that it is not necessary to get question 3 correct in order to get question 4 correct. People sometimes think it means that one cannot have an "item cluster" or "item bundle" like a series of questions linked to the same passage, or a cloze test. This point is not very relevant to developing scale descriptors, but Theunissen discusses how the former can actually be coped with (Theunissen 1987), and Hill has successfully developed a cloze item bank (Hill 1991). Finally, as stated above, the guessing problem is mostly associated with multiple-choice and true/false items, and not particularly relevant to descriptor development.

Developing a Framework Scale with the Rasch Model

Wilson (1989) lists three ways in which the use of the Rasch model can assist in establishing a learning hierarchy apart from the meeting the requirement for objectivity as discussed above:

1. The search for item sets (in our case descriptors) of homogeneous difficulty can inform the reshaping of the definitions of tasks, and

in a test-based model, draw attention to weak operationalisation of the tasks in the test items.

2. Insights can be gained into problems with our theories of learning and instruction, to do with sequencing obviously, but also in terms of the size of the step up from one objective, task or level to the next, the existence of thresholds where the rules of the game change.
3. It can give a framework of reference for discussing the behavioural meaning of different levels of attainment in a learning sequence (i.e. you can develop descriptors).

Rasch is eminently suitable to both establishing a continuum of proficiency for criterion-referencing (Masters 1990: 62–5) and to intra- and international comparisons of achievement as for example in IEA studies (International Studies in Educational Achievement (Masters 1993). Three examples from Australia of relatively simple applications of a Rasch model to the development of scales of language proficiency are as follows.

In the first project, aspects of competence relevant to writing tasks were placed in a hierarchy. Defined assessment scales were developed from them and then writing samples were calibrated to the scales. The scale and samples worked as a "ladder of developing competence," as a framework to guide scoring (Harris et al 1988 cited in Masters 1988a: 296).

In the second project, the results from test tasks written in order to test specific subskills were analysed and placed on a linear scale. Once on a linear scale, the test-tasks were divided into three groups for Level 1, Level 2 and Level 3. Short descriptors defining the subskills tested by each test task were then substituted for the names of the tasks on this scale, giving a hierarchy of descriptors in three levelled groups. Finally the group of discrete task-descriptors for each level were edited into paragraph-length certificate statements (Brown et al 1992).

In the third project, groups of primary teachers collected samples of reading behaviour and described them in short "indicators." Those descriptors commanding a consensus were included in matrixes (type of observation x type of behaviour) on different subskills. An observation survey of classroom reading behaviour was undertaken using rating questionnaires made up of descriptors with a 0–3 rating scale attached. The survey results were analysed and placed on a linear scale giving a hierarchy of descriptors, which was studied for patterns and summarised into bands. The draft scale of bands was then circulated for consultation to experts and teachers, used

in field trials to check that the bands reflected reading behaviour and revised into a first edition (Griffin 1989).

All three examples demonstrate the development of empirically based standard setting for criterion-referenced assessment, exploiting the continuum of proficiency plotted with the aid of the Rasch Rating Scale Model (Masters 1990: 57–9; 66).

Unlike with classical measurement, scale values defined with Rasch are "invariant," constant, objective, *provided that* (a) the data fits the assumptions of the model (Hambleton et al 1991: 23) and (b) that a new group to whom the scale is to be applied can be regarded as another subgroup of the same population. At what point a new group should be regarded as a new population (for whom the scale values would not apply) is a value judgement. In certain cases it should be obvious. For example Spolsky's American orthodox Jews learning Hebrew for ritual purposes (Spolsky 1986: 152–7) could hardly be regarded as the same population as West European schoolchildren and young adults. In cases of doubt, it can be tested empirically, as described in Chapter 6 in relation to adult sector and school sector learners.

The Rasch model is most associated with itembanks (e.g. Pollitt 1989; Henning 1987a) and with the exploitation of itembanks in computer-adaptive testing. The methodology applied in this study is an itembanking one. The items are descriptors rated by teachers rather than gap-fill questions filled in by learners, but the principle remains the same. The primary product of this study is a "descriptor bank," the descriptors being items with known scale values and statistical properties, as with items in an item bank. As with an item bank, a descriptor bank can be extended, since the difficulty of all new items can be calibrated onto the same common scale used in the establishment of the original bank. This development can also be in terms of *persons* rather than or as well as items; new sub-groups (from other sectors, other countries) can be added to a future survey, and one can check the extent to which these new subgroups can be regarded as part of the same population. By the same token, a survey can be repeated after an interval with the *same* population to check whether changes (e.g. changes in teaching methods) have had an effect on the calibration of the items in the bank. The Rasch model "descriptor bank" thus offers a way to give an empirical base to the development of a common framework scale, a means to extend it to more and more groups, and a means to prevent it atrophying and possibly becoming a brake on curriculum development.

There are three possible sources of data for establishing a hierarchy among descriptor elements that could be put together into statements for levels in a common framework and all three involve subjectivity in judgement:

1. Designing/identifying tests targeted at the specific skills and sub-skills involved, and allocating descriptors to levels on the basis of the place in the Rasch hierarchy of the test results.
2. Teacher or user judgements, expert opinion: asking people how far they get their students, or how likely their students would be to be able to perform a certain task.
3. Rating observed behaviour of students.

Targeted Tests

In Switzerland evaluation is based almost entirely on teacher assessment rather than examinations. Language tests only exist at the point of leaving weekly vocational education classes at age 18, and leaving Gymnasium at age 19 or 20. Only the test used at the end of vocational education has a common specification, and none of these operational instruments are validated before use. In any case, there are severe problems in taking test items as operationalisation of particular skills, as Brown et al have done in Australia (Alderson and Lukmani 1989, Alderson 1988, 1990a, 1990b). Summarising a series of studies on native-speaker teacher judgement of items testing reading sub-skills, Alderson concludes:

> i) Judges are unable to agree as to what an item is testing;
> ii) Judges are unable to agree upon the assigning of a particular skill to a particular test item;
> iii) Judges are unable to agree about the level of a particular skill or a particular item;
> iv) There appears to be a lack of relationship between item statistics and what an item is claimed to be testing. (Alderson 1988; 1990b: 436)

In view of the difficulty of writing test items to assess a given aspect of competence encountered by the Canadian DBP project (Harley et al 1990), and in view of the inconsistency of teachers/testers judgements on what items are actually testing (Alderson 1988), there appear to be severe difficulties in using test items targeted at specific skills to collect the data. Brown et al (1992) used this approach, but validated the test items they used

through an independent analysis. But such an approach implies a doctoral thesis per test.

Alderson's findings have been attacked over technical details in the experiment: that the raters had no training in what they were supposed to look for (Weir et al 1990; Bachman 1991; Lumley 1993); that poor items which could be answered without reading the text were not screened out of the experiment (Lumley 1993: 31) and that Alderson talks of an implicational scale, but gave the raters a binary judgement task (Matthews 1990b). However, these points reinforce Alderson's argument that it is no easy matter defining what items really test, that the items certainly did not appear to test what the author had designed them to test (1990: 431) and that validation is necessary before one can state that an item operationalises anything (1995). Pollitt suggests that it is important to collect many specific examples and emphasises that it is not always possible to generalise across tasks in this way (Pollitt 1993). And one should not forget that matching test items to scale descriptors is a great social responsibility: "When notions like *the ability to extract meaning* become operationalised as scores on, for example, reading tests, a child who fails is then labelled as one who is *unable to extract meaning*. Similarly, when the *cognitive aspects* of language are tested in terms of say, being able to produce synonyms, then the child who cannot is branded as *lacking in the cognitive aspects of language development*" (Martin-Jones & Romaine 1986: 29).

Expert Opinion

The second approach also seems highly questionable because expert opinion has been shown to be very subjective and erratic in general (Meyer and Booker 1991) and particularly in relation to judging the difficulty of items—standard setting (Glass 1978: 248–9). Alderson, for example, found that experienced and inexperienced non-native speaker testers in Sri Lanka were unable to predict the difficulty of test items with any clearly discernible pattern (Alderson 1990a). In a separate study by the US National Foreign Language Center, a hierarchy of difficulty for an Arabic test derived through a Rasch Model analysis of Egyptian native-speaker teachers estimation of difficulty proved to show little discernible relationship to the performance of American students on the test (Lambert personal communication). In another study Alderson found that cut-off scores for an examination varied dramatically depending on whether judges were asked to fix a cut-off mark

for papers, a percentage right/wrong for each question, or an overall cut-off for the examination (Alderson 1990a).

This reinforces Glass's observation that "the specific techniques employed in setting an examination standard may be a more powerful determinant of the standard than any other variable" (Glass 1978: 249). Alderson's overall conclusion is that when we think of problems with judgements in language assessment, we should not just focus on raters. All judgements in the testing process need to be corroborated before they can be accepted as valid (Alderson 1990a, 1990b).

Since the first two methods are problematic, and since it proves to be easier to deal with subjectivity when it is a rating of something observed rather than an abstract impression, this suggests following a behavioural rating strategy.

Rater Judgements

This third method, rating of behaviour, offers the advantage that descriptors can be calibrated directly in relation to learner performance without the interpretation that the other two methods require. There *is* no need to interpret from the specification to the test items to the reporting statement since all three are the same descriptor; there *are* no abstract decontextualised "expert" judgements to be made since teachers are asked to use this set of descriptors to rate their learners at the time of year when they are already focusing on rating their learners for end of year reports. One is still of course left with the problem of the subjective interpretation of each descriptor by each teacher, with problems of inter-rater reliability, rater errors and relative rater severity, but at least one is then dealing with such problems directly, and not with problems at second hand.

Such rating of behaviour can be undertaken in two ways, as was discussed under Behaviourally-based Rating Scales:

- Rating the behaviour of a number of students in a class using a checklist of tasks, quality statements and summary statements. As with Behavioural Observation Scales (BOS) in work evaluation, this would be a mixture of direct observation, retrospective observation, and expected behaviour (Griffin 1989).
- Rating common samples of behaviour representative of the range of levels in the system.

Inter-rater Reliability. The problems of inter-rater reliability (consistency between raters) associated with the above are well known, and a reason that some people have pursued supposedly "objective" strategies like error counts for which there is even a version of the Rasch Model, called the "Poisson count." The problem was recognised at the end of the 19th century. Edgeworth estimated the degree of chance in public examinations to between one third and two thirds (Edgeworth 1890: 563, cited in Linacre 1989: 10) and Ruggles noted that the amount of variance between judges was as great as that between candidates (Ruggles 1911, cited in Linacre 1989: 10). The American National Board of Medical Education dropped subjective assessment after studies demonstrated inter-rater reliability of only 0.25% (Hubbard 1971, reported in Raymond et al 1991). Not much progress has been made in that area in the health professions. Cason and Cason (1984) demonstrated 35% of variance due to how strict the rater was, and only 40% of the variance to ability. A review of inter-rater reliability studies in work evaluation reported some findings in the 0.70s and 0.80s occurring, but states that the majority are around 0.40s to low 0.60s—meaning that estimated variance accounted for by ability is only 20% to 40% (square of the correlation) (Muzzin and Hart 1985, cited in Raymond et al 1991). Borman is reported to have conducted a very carefully controlled experiment with a rigorously constructed scale, well chosen samples, trained raters, lab conditions, and to have achieved only 0.80 or 64% (Linacre 1989: 10).

The general response has been the trend towards defined descriptors and a focus on training. Jason gives typical advice to make the scales as clear as possible by refining the aspects to be rated, refining response categories and training the raters. In this way, Jason claimed that scale reliabilities of 0.86–0.93 were obtainable (Jason 1962, cited in Wolf 1988: 496). The language testing literature appears to reflect both Borman's and Jason's experience. One needs to make a distinction between what is achieved in what might be called "studio conditions" by a small number of often expert raters rating a small number of learners during the development and initial validation of a scale, and the coefficient achieved in large scale operational use. For the latter, Alderson, Clapham and Wall (1995: 132), Cohen (1994: 36) and other language testing handbooks suggest that an inter-reliability coefficient of 0.80 is reasonable to aim for in large-scale operational circumstances for standardised tests. For the Test of Spoken English, an inter-rater coefficient of 0.82 is reported for single raters with 0.90 for averages be-

tween two raters (Educational Testing Service 1995: 9). This appears to reflect Borman.

For smaller scale "studio studies" considerably higher mean inter-rater reliabilities have been reported. Coefficients of 0.91 (Adams 1978); 0.93 (Henning 1992b); 0.93–9 (Dandonoli and Henning 1990) and 0.96–9 (Stansfield and Kenyon 1992) have been reported for FSI/ ILR / ACTFL oral proficiency interviews. Such results are echoed by other "studio studies," for example in relation to a Hebrew rating scale (Shohamy 1981: 0.98) and the Cambridge Assessment of Spoken English (Milanovic et al 1992/6: 0.93). High mean inter-rater reliabilities have, however, also been achieved for day-to-day operational use of very well-defined language testing approaches, notably in relation to ACTFL (Thompson 1995: circa 800 telephone interviews: 0.85–0.90) and the Michigan writing scales (Homburg 1984: mean 0.88; Hamp-Lyons 1990: circa 0.90).

Inter-rater reliability can be expected to be a significant problem in all but the most strictly moderated approaches with trained raters using commonly agreed, defined criteria, and in controlled "lab" experiments. Many writers would not in any case accept that inter-rater reliability is in itself a viable goal, following the argument that the one thing that would be certain if a group of raters agreed entirely would be that the agreed rating was wrong: agreement may be an agreed bias rather than a valid measure (Saal, Downey and Lahey 1980). This point can be taken further, to say that the reliability of a rating scale (however established) says nothing about its validity, since the reliability can just be consistent bias and is not synonymous with rating accuracy (Wherry 1952). "Examiners who agree are likely to be wrong" (Harper and Misra 1976: 260 in Linacre 1989: 21).

Rater Errors. The classic rater errors are halo effect: transferring judgements from a global impression to categories, or between categories, central tendency: not using the top and bottom of the scale or tending to home in on a neutral category on a questionnaire, and variation in severity/leniency. Training in the work performance evaluation field tends to concentrate on these points, that is to say on changing rater behaviour. While there are indications that these problems can be reduced by video workshop training (Cooper 1981: 233 in relation to halo reduction, Ivanovitch 1979 in relation to halo and severity/leniency) and by related diary keeping prior to assessment (Bernadin and Walter 1977 in relation to leniency), "successful" training to change rater behaviour paradoxically does not necessarily in-

crease rater accuracy or inter-rater reliability and can in fact reduce both (Bernandin and Pence 1980, Borman 1979). Finally, the effects of training appear to diminish over time (Ivanovitch 1979, Cooper 1981 citing Warmke and Billings 1979, Bernadin 1978).

These findings may reflect the fact that two of the classic errors, halo effect and severity/leniency, are personal characteristics of the way the rater rates, different routes to the same goal (Einhorn 1974), which are extremely resistant to training (severity: Linacre 1989, halo: Cooper 1981). To be "effective" therefore, training concentrating on these classic rater errors needs to destabilise the rater's natural approach, disorientating the rater in the process, which may be what causes the kind of loss of reliability and accuracy reported (Bernandin and Pence 1980, Borman 1979). Such training aiming to change rater behaviour in relation to "classic" errors may be totally misconceived. After a review of studies bearing upon the relationship between rater accuracy and halo effect Cooper reaches the devastating conclusion that: "…the best available estimates suggest that halo and accuracy share a median of 8% of the variance, but the direction is opposite to the prevailing assumption: that is, higher halo and higher accuracy modestly covaried" (Cooper 1981: 239).

Cason and Cason suggest that three factors determine how a rater rates: (1) the "resolving power:" whether or not someone can make decisions; (2) the "rater reference point" RRP: a pivotal implicit standard; (3) sensitivity (Cason and Cason 1984). For example, a teacher who usually teaches the 4th grade will have a standard, an internalised norm of what a prototypical 4th grader would achieve. English language teachers who frequently take First Certificate students may have RRPs which are very close to each other, having over a period of time internalised the norm for a "pass" at First Certificate. A rater with high sensitivity is a rater who discriminates well in the immediate vicinity of his/her RRP the way many class teachers can rank their students for overall ability. A rater with low sensitivity is one who rates well over the whole continuum, not necessarily any better near his or her RRP. If Cason and Cason's internalised standard (RRP) is holistic, as they imply, this may help explain the complex interaction with category judgements which we call halo effect.

Borman suggests that rather than focusing on classic rating errors, training to improve inter-rater reliability requires a *common nomenclature* like a *frame of reference* for defining effectiveness levels, as well as the obvious focus on standardising observation and agreeing weightings of qualities. He sug-

gests that training to increase accuracy requires that these agreed-upon effectiveness levels and weights attached to different factors should be "correct" and "uncontaminated" by irrelevant aspects, which implies that they should be objective measures of subjective judgements (Borman 1979).

This is very similar to what was being aimed at in this study: the creation of a common framework of reference made up of defined effectiveness/proficiency levels established by the objective scaling of subjective judgements—after the identification and removal or correction of "contaminating" factors. A very significant question in such an enterprise is whether one should try to do anything to *reduce* rater errors in the judgements which become the raw data set, or whether one takes so-called naive raters, with all their errors, and uses a methodology to identify and correct for them. The idea of training raters for such an enterprise is attractive but dangerous. Firstly, the training would be in terms of a standardising interpretation: *What do these descriptors really mean? Are you being too strict?* In this particular case that would have been putting the cart before the horse, since the aim of the study was to find out what the descriptors meant, and how strict different kinds of teachers are when they interpret them. Secondly, the training would be in terms of a standardising procedure and the experience of the limited training which might be feasible could be expected to disorient the teachers and destabilise the accuracy of their judgements (Bernadin and Pence 1980, Borman 1979).

It therefore seemed preferable in this study to take teachers as they were and to apply a methodology which would (a) provide information about variability and error, (b) provide the means to identify and exclude contaminated data and still provide a generalisable result, and (c) take account of differences in the severity of the teachers in judging the proficiency of their learners. The Linacre many-faceted version of the Rasch rating scale model offers such a technique.

The Many-faceted Rasch Model

The Linacre (1989) model is a significant development in that it allows one to add a third measurement facet "judge" to the two conventional facets "item" and "person" in order to estimate the severity of each rater. One can then take the "difficulty" of the rater into account in estimating person ability in the same way that the conventional two-facet Rasch model takes just the difficulty of the questions into account in estimating person ability. The Linacre model gives item free, person free, judge free measurement. It can

provide valuable information about the way in which the rating scale is used by different raters and the way in which it is used in combination with different items, allowing this also to be taken into account in estimating ability and difficulty values. Other facets such as the rating occasion can also be added and the model allows the way in which these different facets interact to be studied, and appropriate adjustments to be made to arrive at a fairer, more accurate judgement. As Linacre sums up: "accurate measurement depends not on finding one "ideal" judge but in discerning the intentions of the actual judges in the way which they have replicated their behaviour in all the ratings each has made" (Linacre 1989: 43). The model can therefore provide information about variability and error, enables one to identify and exclude contaminated data as in any Rasch analysis, and can take not only the severity of the rater, but also the conditions under which he/she is operating into account.

Although the approach is mainly concerned with severity (the judge), it is interesting that this focus on the way facets interact in the rating context also reflects the state of the art on research into the halo effect. Murphy and Anhalt propose that rather than being a rater-related issue as it was interpreted by Cooper:

> "Halo error may reflect a wide variety of influences including the rater, the ratees, and the specific behaviour that is being evaluated at a given point in time." (Murphy and Anhalt 1992: 499)

In the terminology of studies which have used the Linacre model, rater = judge; ratee = person; specific behaviour = item (or item on a task); given point in time = rating occasion.

Like all item response models, the Linacre model uses a linking network of anchor items: items common to different tests or questionnaires which "anchor" them together. It therefore requires a data collection design which gives a linked network. It can cope with incomplete data and is very economical in the way the network is constructed.

Use of the Many-faceted Model in Framework Development

A Rasch model scalar analysis of the way in which descriptors are used to assess learners, and in particular an analysis using the Linacre version of the model outlined above, is able to answer a number of the requirements for a common framework scale which have been identified in this chapter.

Firstly the judgements of people representative of the target user population can form the basis of the analysis so that the usability of the descriptive model and descriptor style for the kind of people concerned is established. Descriptors to which these users relate well (i.e. descriptors they use consistently: descriptors which "fit" the probabilistic expectations of the Rasch model) can then form the basis of the final product (Smith and Kendall's 1963 philosophy). During this process, systematic variation in the way different groups of users interpret different kinds of descriptors can be investigated (Borman 1979, Richterich and Schneider 1992).

Secondly, by working at an item level (each descriptor = an item) one can incorporate elements from existing scales and sets of levels which may later help in the interpretation of the scale constructed. In this way the experience gained in the field can be built upon, but without incorporating false assumptions which systematise the error they are designed to reduce (Landy and Farr 1983) as happens when conventions are copied as clichés from one scale of language proficiency to another (North 1992a: 168) without information about the validity of those descriptors in the new setting (Brindley 1991: 6–8).

Thirdly, by using a measurement model to scale the descriptors one can determine the extent to which descriptors for different content areas can be actually measured on one dimension and can therefore justifiably be put in the same scale. The reliability of the scale, and the number of level strata related to it (the justifiable number of proficiency bands) can also be deduced from the analysis of the way people rate.

Finally by using a measurement model which separates and scales the parameters: item, person, judge, one can provide objective measurement from subjective judgements (Thurstone's requirements), and by defining various descriptive factors (e.g. educational sector, language region) as facets, one can measure the effect of such factors in the measurement context.

Relevant FACETS Studies

A number of studies using the FACETS model have been reported since its appearance in 1989, and the most significant for the current study are summarised below.

Myford assessed the way acting ability by high school students was interpreted by different groups of judges: experts, theatre buffs and novices. She used a 6 point Likert scale (Oppenheim 1966/92: 195–200) in a questionnaire of 36 qualities, each identified by a label (Myford 1991).

Stahl and Lunz, in a series of studies, examined practical medical examinations, checking the consistency of judges ratings (did they keep to their standard) across a number of administrations of the exam (Stahl and Lunz 1991, Stahl, Lunz and Wright 1991).

Stansfield and Kenyon investigated how 3 different groups of a total of 402 teachers (bilingual education teachers, French language and Spanish language teachers) ranked 38 tasks from the ACTFL guidelines by asking them to rate on a 5 point Likert scale whether they thought a teacher of French/Spanish in Texas should have an ability level at which they would be able to perform that task. In a second study, they asked the same teachers to listen to 15–17 audio taped extracts from interviews previously rated onto the ACTFL scale by ACTFL raters and answer YES or NO to the question whether this person had a sufficient level of French/Spanish to teach in a Texas classroom. The answers to the questions were used to determine a hierarchy amongst the tasks and amongst the extracts. The hierarchy discovered coincided very closely with that intended in the ACTFL Guidelines, with the exception of 4 tasks with a high non-linguistic or personal dimension (Stansfield and Kenyon 1992).

Tyndall and Kenyon validated a newly developed holistic rating scale of defined descriptors to be used in the placement test for Georgetown University's ESL program. The analysis endorsed the scale and identified one teacher (who had been away when the staff developed the scale, and also missed training) as "misfitting"—and in need of remedial training (Tyndall and Kenyon 1996).

Each of these studies has some relevance to the current study. Myford's comparison of the interpretations and relative strictness of different kinds of judges suggests the possibility of comparing the judgements of teachers from different educational sectors and pedagogic cultures as has been done in this study. Stahl and Lunz's investigation of consistency on different rating occasions suggests the possibility of comparing assessments arrived at through teacher continuous assessment and in a formal summative assessment context. Stansfield and Kenyon's study suggests the possibility of combining a questionnaire survey designed to calibrate descriptors with a second study to measure the respondents severity through the rating of recorded performances, as has been done in this study. Finally, Tyndall and Kenyon's study points out how to identify assessors who do not share the frame of reference of their peers, and who should therefore (in the case of

this study) be removed from the data or (in the case of their study) re-trained.

Summary on Measurement Issues

In scales of language proficiency and Behaviourally Anchored Rating Scales (BARS) for the evaluation of work performance, persistent problems have been caused by inadequate item analysis and insufficiently rigorous determination of the level of the descriptors—with the effect of systematising error rather than eliminating it. The Rasch rating scale model offers a way of constructing "a criterion-referenced scale of developing proficiency from a set of expert ratings and providing a means of evaluating the extent to which different aspects of performance can validly be combined into a single global measure of proficiency" (Masters 1990: 66).

Another attraction of a Rasch approach is that the item analysis, scale validation, sample calibration etc. can all be part of the same databank, simplifying the development process. Secondly, this databank can be continually expanded in a project spiralling outwards in a series of phases, whilst being able to give a concrete outcome with final calibrations at the conclusion of each defined phase. Such an approach to the development of a common descriptive framework provides an empirical psychometric base to support it. Since the "descriptor bank," the learners who were in a survey and the performance samples used are all calibrated onto the same linear scale, this empirical base means that the framework can be presented in different ways for different purposes. For example, the descriptors can be presented as profile grids, checklists or holistic scales; the range of level or categories, even the cut-off points between levels and categories, could be customised for different groups of learners (e.g. lower secondary with one year of English) for different purposes (e.g. continuous assessment, oral exams, reporting to parents/employers) and still relate back to the same common scale. Since performance samples can be rated onto the same scale in the analysis, the meaning of the descriptors can be made even clearer through training in relation to those samples.

4 Developing a Descriptor Pool

As has been commented, it is customary when undertaking the development of a new scale of language proficiency to use existing scales to which access is available as sources. Interpreted negatively one could say that in this way conventions and clichés get copied from scale to scale without an empirical basis (Brindley 1991: 6–8; North 1992a). Interpreted positively one could say that this is a reflection of the consensus over achievement in foreign language learning at different levels of competence. Since the aim of this study was to create a meta-system, a transparent and coherence *common* framework, it made sense to start from a comprehensive survey of existing scales. Such a survey of scales describing Spoken Interaction and/or Global Proficiency was produced in the context of the Council of Europe project (North 1993a), and provided the starting point for the development of a pool of descriptors for this study. The scales used as sources, reviewed in detail in the survey, were the following. The abbreviations are those used in the classified descriptors in Appendix 3.

Scales of Overall Competence in Spoken Interaction	
Hofmann: Levels of Competence in Oral Communication 1974	**Hof**
Wilkins: Proposals for Level Definitions for a Unit/Credit Scheme: Speaking 1978	**Wilk**
University of London School Examination Board: Certificate of Attainment 1987	**Lon**
Ontario ESL Oral Interaction Assessment Bands 1990	**OTESL**
Finnish Nine Level Scale of Language Proficiency 1993	**Finn**
European Certificate of Attainment in Modern Languages 1993	**EurLon**
Sets of Sub-scales for Competence in Different Contexts	
Trim: Possible Scale for a Unit/Credit Scheme: Social Skills 1978	**Trim**
Eurocentres: European Language Portfolio Mock-up: Interaction Scales 1991	**North**
Association of Language Testers in Europe, Bulletin 3, 1994	**ALTE**
Detailed, Holistic, Rating Scales	
Elviri et al: Oral Expression 1986 (in Van Ek 1986)	**Elviri**

FSI family:	
Foreign Service Institute Absolute Proficiency Ratings 1975	**FSI**
Interagency Language Roundtable Language Skill Level Descriptors 1991	**ILR**
Australian Second Language Proficiency Ratings 1982	**ASLPR**
American Council on the Teaching of Foreign Languages Proficiency Guidelines 1986	**ACTFL**
ELTS family:	
Carroll B.J. and Hall P.J Interview Scale 1985	**C&H**
International English Testing Service (IELTS): Band Descriptors for the Speaking Test 1990	**IELTS**
English Speaking Union (ESU) Framework Project: Speaking 1989	**ESU**
Analytic Rating Scales	
Carroll B.J. Oral Interaction Assessment Scale 1980	**carr**
Hebrew Oral Proficiency Rating Grid 1981	**sho**
Goteborgs Univeritet: Oral Assessment Criteria	**Got**
Fulcher: The Fluency Rating Scale 1993	**Fulch**
Frameworks of Content Specifications and Assessment Criteria for Stages of Attainment	
University of Cambridge/RSA Certificates in Communicative Skills in English 1990	**CCSE**
Royal Society of Arts Modern Languages Examinations: French 1989	**RSA**
English National Curriculum: Modern Languages 1991	**NatC**
Netherlands New Examinations Programme 1992	**Dutch**
Eurocentres Scale of Language Proficiency 1993	**EC**
British Languages Lead Body: National Language Standards 1993	**Llb**

A manageable descriptor pool required organisation by categories and by levels. The way the pool was created was by taking, scale by scale, the contents of the each of the scales for global proficiency and for spoken interaction / speaking listed above and breaking that content up into sentences. Each sentence was then analysed to see what category it seemed to be describing and allocated to that category in the descriptor pool.

Provisional Categories

The categories were arrived at through (a) a consideration of the contents of the scales, (b) reference to theory as discussed in Chapter 2, and (c) reference to the categories under discussion in the Council of Europe Frame-

work authoring group. A table showing the set of categories which was the result of the process was given at the end of Chapter 2.

Provisional Levels

After an attempt at making finer distinctions proved unworkable, descriptors were grouped into the seven broad levels proposed by Wilkins (1978). Since it was not proposed to produce a questionnaire at the level of near-native proficiency, this led in effect to the adoption of 6 levels. It is perhaps not a coincidence that these 6 levels appear to correspond to the "natural levels" proposed by Hargreaves (1992) of which the upper 5 have since been adopted by ALTE (Association of Language Testers in Europe). There does actually appear to be quite a wide consensus on the nature of broader conventional levels. One should not, however, forget that these levels were at this point being used purely in order to organise raw material. The purpose of the study was to calibrate the descriptors, and see in so doing how many levels could actually be justified from the reliability statistics for the data. The levels adopted at this stage were labelled as follows:

Breakthrough corresponding to what Wilkins in his 1978 proposal labelled "*Formulaic Proficiency*" and Trim in the same publication (Trim 1978) "*Introductory.*"

Waystage reflecting the existing Council of Europe content specification.

Threshold reflecting the existing Council of Europe content specification.

Independence described as "*Limited Operational Proficiency*" by Wilkins, and "*adequate response normally encountered*" by Trim. The expression *Independent User* has been adopted by ALTE.

Effectiveness called "*Effective Proficiency*" by Trim, "*Adequate Operational Proficiency*" by Wilkins corresponds approximately to Level 6 on IELTS (International English Language Testing Service) and to the DALF (Diplome Approndi de Langue Française), both standards for students who wish to study at university.

Mastery Trim: "*Comprehensive mastery,*" Wilkins: "*Comprehensive Operational Proficiency*" covers the top examination objective in the ALTE scheme (Cambridge Proficiency).

Provisional collation was undertaken partly on the basis of relationships between the scale "families" and partly through comparison of video/audio recordings (between ESU, Eurocentres, ILR, CCSE) undertaken previously in a Eurocentres context. This provisional classification was then refined though a detailed comparison of the wording during the process of the collation.

Editing

A collation of all the descriptors yielded a total of 1,679 potential descriptors. The next task was to reduce and edit this mass into a pool which could then be refined in the planned series of workshops with teachers which are described later in the chapter.

Thus far only descriptors from scales for speaking, spoken interaction or for "global language proficiency," which tended to also focus on spoken interaction, had been included. At this point, however, scales for writing were investigated to see to what extent the statements made about qualitative aspects of proficiency might also apply to spoken language, particularly spoken production (sustained monologue). At this point therefore, the content on aspects of competence / proficiency for all the scales for writing from scales of proficiency used so far as sources was collated into the pool, plus descriptors from the experimental analytic writing assessment scale developed by Hamp-Lyons and Henning (1991). Finally, since listening to the interlocutor(s) is an integral part of spoken interaction, selected descriptors for listening in interaction, need for interlocutor adjustment and ability to maintain conversation were incorporated from listening scales, particularly the Australian Migrant Education Program Scale (AMES). This gave a total of 2047 potential descriptors.

In the editing process that followed, the aim was to remove repetition and reduce the number of descriptors to a manageable number, but also to try and produce positively worded "stand alone" statements that could be independently calibrated. Trim (1978: 33–34) criticised scales of language proficiency for including much normative, negative wording, which prevented the descriptors in them from being used to express objectives. Skehan criticised scale descriptors for only having meaning in relation to other sentences in the same paragraph, or other statements in the scale. He considered that assessment in relation to scales of proficiency produced like ELTS could not be regarded as true criterion-referenced assessment because the content of the scale could not be represented as a series of cri-

terion statements to which a Yes/No answer could be given (Skehan 1984: 217). The aim of the editing process was to try and reformulate the content of scales so as to provide a pool of such criterion descriptors which could then be put through a qualitative validation process with teachers before being used as criteria for rating students with Yes/No judgements, differentiated by performance conditions.

After the editing process was complete it became apparent that the coverage of Strategic Competence and of Socio-cultural Competence was particularly weak.

Strategic Competence was equated almost entirely with compensatory and repair strategies, and there was not a single reference to compensatory strategies at *Waystage,* where they might be thought to be most crucial (Van Ek in North et al 1992: 20), except one negative one from Wilkins. Interaction Strategies were also poorly dealt with, the British National Language Standards and Eurocentres providing the only descriptors on Co-operative Strategies, the former on Cognitive Strategies (Cooperating: ideational) and the latter on Collaborative Strategies (Cooperating: interpersonal), to use Barnes & Todd's classic (1977) distinction. Asking for Clarification (discourse "challenge:" Burton 1980) is only dealt with in detail by the RSA and then only at *Effectiveness* Level (RSA3). Some 60 new descriptors were therefore written to cover various aspects of strategic competence.

Socio-cultural Competence was also minimally dealt with in terms of levels of politeness and formality and, occasionally, register flexibility. Unfortunately it was not possible in the short time available between preparing the descriptor pool and the start of the workshops to work on both these areas; a choice had to be made, and since a framework had by this stage been worked out for strategic competence, that area was chosen.

Pre-testing Workshops with Teachers

Having now got a pool of approximately 1,000 quite reasonable looking descriptors, the next step was to check that the descriptive categories used made sense to teachers, that the descriptors described the category in which they had been put, that teachers found the wording of the descriptors transparent and helpful, and that the descriptor pool was relevant to the educational sectors concerned. Secondly, a pool of just under 1,000 descriptors was all very well, but it was clear that with the 6, 7 or 8 questionnaires planned for the main data collection, it would be impossible to use

more than between a quarter and a third of them, so some principled means of sorting out the best ones was necessary.

Eleven workshops were undertaken in March and April 1994 with teachers from all four of the educational sectors concerned: lower secondary, upper secondary, vocational education for apprentices, adult education evening classes and for the two main language regions: German-speaking and French-speaking. 10 workshops concentrated on descriptive categories, and the last, a large workshop with 25 teachers, on sorting the descriptors into levels. Colleagues in the project undertook another 7 workshops in the same period to the same format. Two techniques were used at all workshops. The first technique concerned discussing a recorded video sample at a level relevant to the teachers concerned and the second technique was a category sorting task with descriptors.

Investigating the Proficiency Categories used by Teachers

The first technique was simplified from one which was used by Pollitt and Murray (1993). In their experiment, raters were asked to rate which of a series of pairs of learners talking to each other was better. Since each learner talked to each of the other learners in one recording, they were able to apply Thurstone's (1959) law of comparative judgements to scale all the learners. After each judgement, in an application of Kelly's theory of personal constructs (Kelly 1955) the raters were asked to discuss how they felt the two performances were similar and how they were different. According to Kelly people "actively build up a system of constructs for making sense of the world which is constantly undergoing modification as they experience new events or different outcomes for familiar events" and "each individual has their own repertory of constructs, and repertory grid analysis is a procedure designed to elicit from an individual how they view the world" (Pollitt and Murray 1993: 4). Pollitt and Murray were thus able to see "how their raters viewed the world" in relation to performances which they were able to independently scale, and thus they were able to deduce what aspects of proficiency might be salient to raters at different proficiency levels.

In the simplified version of the technique applied at 15 workshops in this project, a small "bank" of video recordings of interactions between pairs of young adults and as far as possible French and German-speaking learners which had already been calibrated in other contexts was put together. The seven videos shown in Table 4.1 were used:

At each workshop, the group of teachers were presented with a performance thought to represent a performance at approximately the level they taught. Immediately after viewing the video, small groups of 2–5 teachers were asked to discuss the performances of the two learners in the recording to the following instructions:

> *Which of the two people in the video is better?*
> *Say what you think and justify your opinion by referring to aspects of the two performances.*

Table 4.1: Video Recordings used in Workshops

Targeted Level	Sex	Mother Tongue	Topic/Activity
Waystage	2 F	G & F	Holidays - conversation
Waystage	1 M	F	Holidays - presenting with a map
Threshold	M & F	F	Planning birthday surprise
Independence	Mixed	F	Simulated teachers' meeting
Ind/Effect.	Mixed	G & F	Simulated teachers' meeting
Effectiveness	2 F	Sp & It	Car park site - conversation
Effectiveness	M & F	G & ?	Learning English - conversation

These discussions were recorded and transcribed in note form. The transcriptions were then analysed using as far as possible the same categories that had been used to analyse the source scales.

The aim of this exercise was threefold:

1. To check that the categories which teachers used as a metalanguage were in fact represented in the descriptive structure used to organise the descriptor pool.
2. To check the extent to which what they were actually saying could be said using descriptors available in the pool.
3. To see if any of the teachers' descriptions of aspects of performances could be edited into descriptors saying something which was not already said by descriptors in the pool.

All the Competence categories discussed in Chapter 2 were used by teachers in the workshops. This suggested, apart from the fact that the video recordings had been well chosen, that teachers did in fact use these categories to discuss performance and that therefore basing the descriptor pool on them was justified. On the other hand, only two categories were used by teachers which were not represented in the descriptor pool. These were: risk-taking as a strategy, and language awareness. After some discussion in the project team, it was decided that the former, reflecting Rubin's (1975) "Good Language Learner" was not always positive, and that both categories concerned learning strategies which, whilst of interest to teachers, would not be appropriate for inclusion in definitions of competence/proficiency at a series of levels.

Pre-testing Descriptors and Categories with Sorting Tasks

The second technique was based on that used by Smith and Kendall (1963) in the development of arguably the first defined assessment scale with calibrated descriptors.

First of all, the descriptors in the pool were edited where necessary to avoid obvious disclosure of a category. For example the word "meeting" was replaced with "discussion" in all the descriptors on Meetings. Pairs of teachers were given a pile of 60–90 descriptors cut up into confetti-like strips of paper and asked to sort them into 3–4 labelled piles, which represented potential categories of description. The categories would be related, for example: Fluency, Flexibility, Coherence (all Pragmatic Competence). At least two and generally four pairs of teachers (either at the same or succeeding workshops) sorted each set of descriptors. Where uncertainty over categories became apparent, up to as many as ten pairs were used.

Labels were written on envelopes into which the descriptors were later put. An extra envelope marked "Unclear / Unhelpful" was also provided. Teachers were asked to put into this pile any descriptors for which they could not decide the category, and any descriptors which they found unclear, verbose or otherwise unhelpful.

During the first workshop the technique was developed further into a number of steps, applied at all subsequent workshops, though there was not always time for all groups to complete all the steps:

1. Sort the descriptors into the piles—including the discard pile (Unclear/Unhelpful). This was sometimes done individually.
2. Go through each pile together, check that both partners in the pair agree that the descriptors are in the right place; put a tick on those found particularly clear, transparent and useful.
3. Go through the ticked descriptors a third time and circle the ticks on descriptors which were not only transparent and useful but also relevant to the range of achievement in the sector concerned. This last technique was used mainly in connection with descriptors on communicative activities to establish curriculum relevance.

All results were coded and item histories developed as the same set of items was tried out in different workshops. The codes used were those shown in Table 4.2.

Table 4.2: Codes for Sorting Tasks

Code	Meaning	Explanation
$0	Correct category:	It landed in the right envelope
$1	Unclear / Unhelpful:	It was rejected
$2	Wrong Category	It landed in the wrong envelope, the error being noted
$7	Wrong category—but particularly clear and useful	Ticked, but in wrong envelope
$8	Particularly clear and useful—but where does it go?	Ticked, but in Unclear / Unhelpful
$9	Correct category—and particularly clear and useful	Ticked, and in correct envelope
$10	Particularly clear and useful, particularly relevant to my sector	Ticked and circled

The best descriptors were those coded $9 or $0/$10. Codes $7 and $8 indicated that the descriptors were useful, but that there was a problem with the categories.

During the workshops, once they had done the initial sorting, teachers were encouraged to reword or split descriptors which they thought could be

improved. The general tendency in the workshops as a whole was for teachers to reject or alter descriptors which were:

- longer than about 20 words;
- saying what the learner cannot do;
- double-barrelled, being linked by an *and* or *but* or contrasting (can do x but cannot do y), the style used in the ELTDU scale (ELTDU 1976) and the ALTE level descriptors (ALTE 1994);
- dense, verbose or jargony (UK Language Lead Body National Language Standards suffered here);
- very classroom oriented with specific context-bound examples (UK National Curriculum suffered here).

The sorting tasks were used, then, to (a) identify the clearest descriptors, (b) get feedback to reformulate where necessary and (c) to check the practicability of the categories being employed. A good half of the descriptors were identified in this way as poorly conceived or poorly worded. There were also some changes to the categories. One example is that it proved impossible to establish a firm distinction between:

- Casual Conversation
- Casual (Café) Discussion
- Collaborative Problem-Solving Discussion
- Negotiating (with the word "negotiate" removed)
- Formal Meetings (with the word "meeting" removed)

Therefore Casual (Café) Discussion and Collaborative Problem-solving Discussion were merged into one single category Informal Discussion, in contrast to Formal Discussion (Meetings).

Checking Provisional Levels

In the majority of the workshops, the tasks were designed to check the quality and relevance of both the categories and the descriptors in the pool themselves. It was, however, also necessary to confirm the organisation of items into the provisional levels. Therefore in one large, final workshop with 25 teachers of English to adults, all the then 400 or so surviving descriptors were sorted into piles by level (Thurstone 1928b) in order to exclude any which were interpreted inconsistently. Two pairs of teachers sorted each category first into three piles, low - middle - high, and then,

when feasible, into two subdivisions of each broad level to give 6 bands, as used in the original collation of items from the proficiency scales. Descriptors were again coded, this time as follows:

\$0	Right level
\$1	"I can't judge the level of this, it could apply to any level"
\$4	Placed apparently randomly: ambiguous as to level
\$5	Placed consistently too high
\$6	Placed consistently too low

There were few surprises with descriptors on the Competence side, as one might expect. As regards Communicative Activities however, there were a number of instances where it appeared that the difficulty of particular tasks were exaggerated in the source scales. For example the following two descriptors on Service Encounters, placed at about *Threshold Level* in the source scales, were put as "Low" by both groups of teachers. This confirmed the high number of \$10 codes (relevant to my sector) given to them by lower secondary teachers, so in the survey, they were used on the two lowest questionnaires aimed at *Breakthrough* and *Waystage*. In the Rasch analysis, both in fact came out calibrated at the level identified as *Waystage*:

Can make simple transactions in shops, post offices or banks.
Can ask for and provide everyday goods and services.

There are on the other hand a few cases where the original authors' opinion rather than that of the workshop teachers is confirmed by the calibration in the survey. One example is *Can agree and disagree politely* from the London 3rd level (*Threshold*) which was put at "Low" by the workshop teachers but which was calibrated in the Rasch analysis at what has been identified as *Threshold Level*. There are also a couple of examples in which a higher interpretation by the workshop teachers was "proved right", and the authors "proved wrong" as with *Can narrate a story* from the London 2nd Level (*Waystage*) put by the teachers at "Middle" and by the calibration at *Threshold*.

This final workshop was used not so much to exclude items as to confirm on which questionnaires they should appear if we wanted to use them. The selection of the items used in the questionnaire survey was undertaken in relation to the supposed level of the learners of those teachers who vol-

unteered to take part (which excluded most of the very high level items), ensuring balanced coverage across all questionnaires.