5 Data Collection and Correction

Rasch modelling, like all forms of item response theory assessment, creates a common scale for a measurement framework. The values of the ability estimates of all persons can be related to each other and to the difficulty value of the items; the difficulty values of all items can be related to each other and also to the ability values for the persons. To do this, the model requires "connectedness" in the data. That is to say the data must either be a full data set (each item being applied to each learner) or it must be linked through what are called "anchor items" or "anchor persons." This study concerned the subjective assessment of learners by their teachers, and the analysis was to be undertaken with the many-faceted version of the Rasch rating scale model (Linacre 1989) which adds "judge" (rater) as a third facet to the conventional facets "item" (descriptor) and "person" (learner). Therefore linking between judges (teachers) was also necessary. Ideally, this would be undertaken by having several teachers rate the same students in a complex matrix design, but this was hardly feasible with 100 teachers and 1,000 learners at the end of the academic year. Therefore, following Linacre's advice, a rating conference was used to achieve linking between the teachers themselves, and between the teachers and the questionnaires which they had not actually used and so to create a common measurement framework.

The descriptors which had survived the pre-testing in the workshops were used to construct 7 overlapping questionnaires of 50 items each. Between them the questionnaires covered a wide range of language proficiency with a total of 280 descriptors. Descriptors were assigned to questionnaires on the basis of (a) their location on the set of 6 provisional levels used to structure the descriptor pool, (b) teachers' statements about the relevance of particular descriptors to their educational sectors (code $10) and (c) confirmation of the approximate difficulty of the descriptors through the sorting tasks by level undertaken in the final larger workshop.

The descriptors were grouped by content under the following subheadings:

- Spoken Tasks: between 18 (top level) and 24 to 27 descriptors per questionnaire
- Comprehension: 4 or 5 descriptors (which sometimes included Reception Strategies)
- Interaction Strategies: 6 to 8 descriptors (which also included Production Strategies)
- Qualities of Spoken Performance: 8 to 10 descriptors (low levels) 12 to 13 (middle) or 16 (top)
- Writing Tasks: 2 to 4 descriptors, with 5 for the top questionnaire

The Writing Tasks were included really just to see what would happen. Would they "fit" with Speaking, or would the analysis demonstrate a lack of unidimensionality?

Each questionnaire was identified by a letter as shown in Table 5.1. Teachers were not told which of the seven in the series they had.

Two questionnaires were produced at approximately the level of Waystage and of Threshold for two reasons. Firstly the number of learners for whom such questionnaires would be suitable, and secondly the sheer number of descriptors available for those levels. The two questionnaires were not exactly parallel because W1 was anchored down to B, and thus could be expected to be "easier" than W2 which was anchored up to T1. Similarly T1 could be expected to be easier than T2, which was anchored up to I. A sample questionnaire is given as Appendix 1.

Connecting Questionnaires

When a Rasch methodology is to be used to calibrate items onto a continuum, the tests (or questionnaires) are linked or "anchored" through either common persons or common items in order to provide sufficient "connectedness." According to Woods and Baker (1985: 129) there is no hard and fast rule over how many items should perform as "anchor items" linking tests; they suggest that 3–10 should suffice in most situations. In this study anchor items comprised 20–25% of the total number of items on each form as recommended by Hambleton, Swaminathan and Rogers (1991). 50-item questionnaire forms were linked by 10–15 items. The anchor items were selected from the items which teachers in the workshops

had correctly classified and ticked, items which were felt to be (a) clearly focused on a particular aspect and (b) transparent and useful.

Table 5.1: Questionnaires for Data Collection

Code	Name	Target Population
B	Breakthrough	Learners with up to 100 hours, which included a large number of secondary school children taking English as a option in their last school year.
W1 W2	Waystage	Aimed roughly at the Council of Europe specification of that name; for elementary learners, including the majority of secondary school learners with 2 years of English.
T1 T2	Threshold	Aimed roughly at the level of the Council of Europe specification, although T2 had a considerable number of items pitched at around the level of a Cambridge First Certificate bare pass. Intended for Berufsschule (apprentices) and other intermediate learners, including younger Gymnasium learners.
I	Independence	Upper intermediate, a level required for new office recruits by Swiss employers, a level which should be reached by the Matura/Maturité at the end of Gymnasium. Intended for FCE classes and the majority of Gymnasium learners.
E	Effectiveness	An advanced level intended for classes studying for the Cambridge Advanced and Cambridge Proficiency classes and for learners in certain Gymnasia known to reach a high standard in English

But even so, there are still choices to be made as to how linkage can best be established. There seem to be three ways in which sufficient connectedness in data can be provided for a Rasch analysis: through horizontal equating; through vertical equating, and through a matrix design. Each will be examined in turn.

Horizontal Equating

Horizontal equating is when two parallel test forms covering approximately the same range of proficiency are linked to create a common item bank. George Rasch originally developed the model in order to be able to provide alternative intelligence tests to assess the ability of Danish army conscripts, thus circumventing the usual problems of test security yet having tests which would report onto the same scale (Wright 1991: 169). The example given by Wright and Stone (1979: 96) on how to connect two tests in what has become the classic Rasch textbook "Best Test Design" similarly concerns horizontal equating. Many applications in language testing relate, like Rasch's original project, to placement tests (e.g. Henning, Hudson & Turner 1985; Blais and Laurier 1993; Chen and Henning 1985). A basic problem with Rasch, however, is that, as will be seen in discussing the analysis, it performs best when the difference between the ability of the learners and the difficulty of the items is not too great. That is to say it is advantageous to get items (in this case descriptors) tried out with learners at approximately the same level as the items. If the difference is too great then the difficulty and ability estimates (calibrations) which are produced become distorted. Horizontal equating is thus less suitable for data collection across a wide range of proficiency because many of the items would get very high scores (nearly everyone can do them) or very low scores (hardly anyone can do them).

Vertical Equating

When developing an item bank covering a broad range of proficiency levels, the alternative method of vertical equating set out in classic form by Woods and Baker (1985) is generally used. Vertical equating requires an overlapping chain of tests targeted at successive levels, linked by the "anchor items." These will be among the easier items on the higher form and among the more difficult items on the lower form. The difficulty of the anchor items on adjacent forms is first calculated only in the context of each form individually, and then the two sets of difficulty values are compared. The average difference for the group of anchors between the difficulty on Form A and the difficulty on Form B is taken to be the difference of difficulty between the two forms as a whole.

Matrix Design

A third possibility is what Griffin (1990a) describes as a matrix design. There are a large number of questionnaires with a large area of overlap and each teacher / rater uses a different questionnaire for each learner, so that each teacher/rater rates every descriptor at least once. In an ideal matrix design each rater not only uses each descriptor or item at some point, but also rates each learner at some point. This method can be feasible in the organisation of a judge-rated examination in a single institution; the more complete the matrix, the more precise the calibration (Lunz, Wright and Linacre 1990).

A matrix implies a circular linking between questionnaires, which implies that nearly all the descriptors must be relevant for nearly all the subjects. Griffin's (1989, 1990a) primary reading survey (the nearest parallel to this study) was able to use a matrix design since it concerned one educational sector which tends to have mixed ability classes within a restricted range of competence. All descriptors were relevant for all teachers, if not necessarily for all learners. Many language testing Rasch applications also relate to groups of students whose proficiency covers a limited range on the full proficiency spectrum which might make a matrix design feasible. For example Adams et al (1987: 15) were working with 270 learners with ratings from 0 to 1+ (approximately *Threshold*) on the ASLPR. Henning (1984: 125) with students with scores on TOEFL up to less than 500 (approximately *Independence*). Madsen (1986: 7) with beginners to intermediate level students. Such restricted ranges of level increase the feasibility of administering the same test/questionnaire to students at the different levels.

Several features of the Swiss project, however, made a matrix design less suitable and led to the choice of a vertical-equating approach. Firstly the Swiss survey was covering a wide range of proficiency from effectively "zero" up to the upper range of the level represented by the Cambridge Proficiency examination, a very advanced level of mastery probably over 650 on TOEFL. Secondly, the data had to be collected during a two-three week period at the end of the academic year when the teachers, all volunteers, were already under a heavy work load. Giving them 10 different questionnaires to rate their 10 learners seemed impractical. Instead all learners in the same class were rated on the same questionnaire. Most teachers did teach at different levels, or in different sectors, and where this was the case those teachers were given different questionnaires for two groups of 5 learners in each level/sector.

It was also impossible to use a matrix design to link the teachers in order to take account of rater severity with FACETS (Linacre 1989–94) since it was impractical for teachers to rate each others' classes. Even if this had been feasible, the instrument, a Behavioural Observation Scale, required a profound familiarity with the subjects gained over a period of time, such as that possessed by teachers about their students at the end of a school year. They could not have acquired such a detailed knowledge of each others' classes, even had they observed the learners for, say, 5–10 hours teaching.

Rating Conference

The question of rater severity was therefore addressed through a one day rating conference in which all participating teachers took part within 3 weeks of rating their students. At the conference, each teacher rated pairs of students on 11 video recordings. These recordings showed selected learners rated by their class teachers in the survey, who were roughly representative of the proficiency levels, sectors and language regions concerned.

Rating Procedure

Ten videos showed pairs and one showed a threesome giving a total of 23 video subjects. An overview tape was prepared which gave a one minute preview of each video, in approximate rank order of difficulty from beginners to a comprehensive idiomatic mastery (*Very* advanced: above Cambridge Proficiency). At the beginning of the rating conference, this preview tape was played as a "taster:" an orientation device intended to give a common reference framework whilst rating each video. In the rating sessions, the videos were then presented in an almost random order and viewed once only.

A mini-questionnaire such as the example given as Appendix 2 was prepared for each video. Teachers were instructed to look at the mini-questionnaire when they got it, and to refer to it during the recording. They were told to make a provisional mark "/" which they could at the end confirm with the second arm of an "X" or cross out and replace. This combination of assessment during the performance and conscious confirmation at the end was based on experience with the development of the Eurocentres oral assessment approach (North 1986, 1991, 1993b) and is set out in the Table 5.2.

Performance Samples

Since the teachers in the survey who had volunteered to record videos were thought unlikely to have had experience of any forms of oral assessment other than interviewing, a standard format was laid down designed to ensure.

Table 5.2: Procedure for Rating Performances on Video

Stage	Action	Function
Before video	Read descriptors	Familiarisation with descriptors and their interaction with the rating scale
During video	1. Refer to descriptors 2. Make provisional mark /	Compare video performance to described behaviour Record "live" impression whilst viewing
After video	1. Review provisional marks, rereading descriptors	Check consistency e.g. in use of rating scale, in rating each person
	2. Confirm or cross out and replace provis-ional mark with X	Make a considered judgement in relation to (a) impression during performance and (b) detached perspective now, re-reading descriptor(s)

Comparability. Since the teachers came from a range of educational sectors and pedagogic cultures, a standard format which would structure the kind of performance was necessary. On the other hand some leeway was necessary for choice of topic by teacher and student.

Autonomy. A video of the learners, not of the teachers was required. Clear and prescriptive instructions limited teacher intervention to setting up the performance and managing the phases. One wanted to avoid learners "reeling and writhing" (Van Lier 1989) and jumping through hoops put in front of them by their teachers.

Interlanguage styles. Tarone (1983) and Ellis (1986) have posited a range of interlanguage styles spanning careful to vernacular and prepared to un-

prepared. Skehan (1987) has proposed that such styles should be taken into account in oral assessment. Secondly, questionnaire descriptors covered both spoken interaction and spoken production (sustained monologue), so one wanted if possible to include spontaneous interaction and semi-prepared monologues in all performance samples.

In order to ensure comparable, autonomous performances which spanned the unprepared / prepared, and interaction / production dichotomies, teachers who volunteered to record their learners were sent model materials and strict guidelines on how to organise the activity. The teacher role: setting up and passing the turn to the second learner for the monologues but otherwise withdrawing to the sidelines, was explicitly stated. Teachers were asked to keep to the overall structure presented in the guidelines, and to use their discretion as regards the topics and cards, exploiting opportunities from their own teaching. All teachers stuck to the decreed format except in one case, in which learners interviewed each other.

The guidelines gave suggestions for topics for Phase 1 (Production) and Phase 2 (Interaction), concentrating on everyday topics like home, holidays, people, pets etc. for elementary and intermediate learners. A set of discussion cards was provided, with the idea that the learners would select a card they wanted to talk about, discuss it as long as they felt like, and then move on to the next topic. For elementary and intermediate levels, the cards were on the same everyday topics, but for more advanced, more academic learners, a set of cards on controversial issues (after Shohamy, Reves and Bejarano 1986) was given. The suggestions to the teachers of elementary classes were as follows:

PHASE 1 Possible Topics for Description Monologue

People:	A friend:	- appearance
		- where do they come from?
		- what do they do?
		- why do you like them?
		- how did you meet them?
Places:	Your home:	- how big is it?
		- how old is it?
		- where is it?
		- what do you like about it?
		- what do you do not like about it?

Pets:	Your pet:	- what's he/she called? - what is he/she? - where and when did you get him/her? - what kinds of things does he/she eat? - what do you like about him/her? - tell a story about something he/she did!

PHASE 2 Spontaneous, unprepared discussion
Tell them there are three topics they can talk about.
Give them the card on the same theme as the Phase 1 description
Tell them they can move onto the other topics if they want to, and put the cards face down beside them.

Homes:	What makes an ideal home? What is important? - where it is? (town/country, buses & trains, shops) - how big it is, the whole place, the rooms? - old/modern? - garden? - modern bathroom?
Holidays:	What kind of holidays are best? Why? - holidays on the beach - adventure holidays, trekking etc. - walking in the mountains
Pets:	What kinds of pets are best? Why? - cats? - dogs?

Altogether about 45 or 50 video performances were collected, from which the eleven shown in Table 5.3 were finally used, with two more (Nos 6 and 12) in reserve. As can be seen an effort was made to have samples from the different language regions and educational sectors. Care was taken to try and get a sample from the adult and appropriate school sector for each level. This was successful at all levels other than upper intermediate, which was represented by just one pair of Gymnasium learners. The performance samples presented in each of the 11 selected videos were as presented in Table 5.4.

The first video, No. 8 from the Ticino, the main Italian-speaking Swiss region, followed an interview format.

Table 5.3: Video Recordings used for Rating

Video	Length	Educ. Sector	Lang. Region	Mother Tongue	Level
Video 1	12.00	Lower Secondary	German	German Spanish	Beginner
Video 2	5.00+	Adult	French	French	Beginner
Video 3	6.10	Adult	German	German	Elementary
Video 4	5.40	Adult	German	German	Intermediate
Video 5	13.00	Vocational	French	French	Intermediate
Video 6	7.45	Adult	French	French	Intermediate
Video 8	10.00	Gymnasium	Italian	Italian	Intermediate
Video 9	6.50	Gymnasium	French	French	Upper Inter
Video 10	12.00	Gymnasium	French	French / German	Advanced
Video 11	13.00	Adult	German	German	Advanced
Video 13	7.45	Adult (University)	French	French / German	Mastery

The interviewer had an English parent, and teachers were asked to focus on the interviewee, Marina.

Table 5.4: Performances on Video Recordings used for Rating

	Production			Interaction		
	Activity	Speaker	Time	Activity	Speakers	Time
8	Interview about a holiday in Istria, describing Istria	Marina (Ital)	*See Total*	Marina's holiday; problems of Italians originating from Istria	Manuela asking and following up	Total 10.00

Table 5.4 (cont.): Performances on Video Recordings

	Production			Interaction		
	Activity	Speaker	Time	Activity	Speakers	Time
1	Pictures of People	Lorenza (Sp)	2.15	Where you live	Both	3.00
	Pictures of People	Nicole (Ger)	2.15	Holidays	Both	2.00
				Pets	Both	2.30
2	Last Weekend	Micheline (Fr)	3.10	Last Holiday	Both	5.00
	Last Weekend	Arlette (Fr)	1.50			
3	Last Holiday	Gertrude (Ger)	1.15	Ideal Holi-day	Both	2.40
	Home	Marcel (Ger)	1.30			
4	Home	Rosemarie (Ger)	1.50	Ideal Pet	Both	2.20
	Home	Renate (Ger)	1.20			
5	Last Holiday	Pascal (Fr)	2.00	Ideal Home	Mainly Marlene	2.15
	Home	Marlene (Fr)	3.00	Ideal Holi-day	Both	3.00
				Ideal Pet	Both	2.00
6	Holiday; learnt juggling	Florence (Fr)	2.20	Juggling	Both	2.00
	Home	Thérèse (Fr)	2.00	Thérèse's Home	Both	1.15
9	Describing a book	Christian (Fr)		Q & A about the book	Both	3.20
	Describing a book	Virginie (Fr)		Q & A about the book	Both	3.30
10	Describing a film	Nils (Fr)	3.30	Card: Pets	Both	3.00
	Describing a book	Sibylle (Ger)	2.30	Card: If women ruled the world.	Both	2.45
11	Meeting a GI on the way to Vietnam	Anne Marie (Ger)		Q & A about it	Both	

Table 5.4 (cont.): Performances on Video Recordings

	Production			Interaction		
	Activity	Speaker	Time	Activity	Speakers	Time
11	Meeting Auschwitz survivor: Film: Schindler's List	Eva (Ger)		Q & A about it	Both	
	Film: Philadelphia: Aids	Doris (Ger)		Q & A about it	Both	13 min total
13	Home	Beate (Ger)	2.15	The Life of an Artist	Both	3.30
	Holiday learning sculpture	Yvonne (Fr)	2.00			

Nevertheless the video with Marina was used as it was the only usable video showing Italian-speakers, and it was important politically to use an Italian-speaking video. This video was the exception in the almost random order, being presented first.

Conference Rating Instrument

The purpose of the rating conference was to provide estimates of rater severity in order to take this into account in estimating the abilities of the learners in the questionnaire survey. To do this it was necessary to have a rating instrument at the rating conference to link the teachers to the questionnaire(s). There were two different forms this instrument could take, reflecting the distinction between Behavioural Observation Scales (BOS: checklist/questionnaires as used in the main survey) and Behaviour Summary Scales (BSS: a profiling grid of levels by categories as used in Cambridge EFL exams and in the Eurocentres assessment approach (North 1991, 1993b).

BSS: The original idea had been to use a rating grid made up of descriptors from the questionnaire survey. Four or five aspects like Range, Accuracy, Pronunciation, Fluency, Interaction would be used to rate performances onto the 6 provisional levels used to collate the source scales. In other

words, the original idea had been for teachers at the conference to rate performance onto *levels* in relation to *categories*. The advantage of such an approach would have been that all the performance samples would have been rated onto the same rating instrument, which would have provided a common point of reference. The conference would have been rather like a moderating meeting for oral examiners.

The disadvantage, however, would have been that subjective decisions would have been made by the author as to which level a descriptor should be used at, since the descriptors had not in fact yet been calibrated. This would have introduced a systematic error into the analysis, and therefore the idea was dropped.

BOS: The alternative was to develop mini-questionnaires for each video using descriptors which related to the performances in the recordings and which were also used on the survey questionnaires at the same level as the recording concerned. The advantage of this approach was that there would be an unambiguous anchor to the main survey as exactly the same items would be being used in conjunction with the same rating scale.

Therefore a series of overlapping "mini-questionnaires" of 5–7 items linked by common items was provided for each video. Descriptors were selected from items used in the survey at that level which had been identified in the workshops as particularly clear and comprehensible. Between 5 and 7 items were chosen because, as discussed in the next section, "Seven plus or minus Two" (Miller 1956) summarises the experience of psychometricians on people's ability to handle categories.

Two or three items described the task being undertaken, one the Production task (e.g. *Can describe pets and possessions*), one the Interaction task (e.g. *Can ask and answer questions about each other, where they live, people they know and things they have*). Descriptors were presented in separate columns for each speaker since both learners in the video did not necessarily have the same task descriptors, since their chosen description task might be different. The task descriptors were followed by about 4 items on aspects of the performance, generally covering Range, Accuracy, Fluency and Pronunciation. A few descriptors for Cooperating Strategies and Compensating Strategies were also scattered through the 11 mini-questionnaires. An example of a video mini-questionnaire is given as Appendix 2.

Rating Scale

One of the issues with which applied psychologists working in behavioural scales for work evaluation have long been concerned with, and which language testers have more recently started debating, is the question of how many levels or steps are desirable in a rating scale. Empirical research in the 1950s suggests that maximum reliability is reached with 5 steps, that this reliability remains constant up to 9 steps, and tails off with either 3 steps or 11 steps. (Bendig 1953, 1954a, 1954b quoted in Landy and Farr 1983). An almost identical conclusion was reached by Lissitz and Green after a series of laboratory studies reclassifying data (Lissitz and Green 1975). Matell and Jacoby reported stable reliability from 2 to 19 categories (Matell and Jacoby 1971), but other studies suggest no increase in reliability above 6 categories (McKelvie 1978). Miller 1956 summarised findings with a rule of thumb "Seven, plus or minus Two" pointing out that psychologists even then had long been using 7 point scales on intuitive grounds (Miller 1956). McKelvie (1978) concludes by recommending 5 or 6 bands.

Following the conclusions of Bendig, McKelvie and Miller the same 5 step scale (0–4) rating scale was attached to each descriptor on both the survey and the conference questionnaires. Five steps rather than, say, 4 steps were chosen because a 5th gives extra precision (Wright and Masters 1982: 136) and because the reservations one commonly hears about rating scales with "middle categories" relate to the *Not Sure* category on attitude questionnaires which, by allowing evasion, may attract a disproportionate number of ratings (Andrich and Masters 1988: 302). On the scale below, based on performance conditions, the middle category is in fact the major decision, and the 0 and 4 really represent extremes.

0 This describes a level which is definitely *beyond* his/her capabilities. Could *not* be expected to perform like this.

1 Could be expected to perform like this provided that circumstances are favourable, for example if he/she has some time to think about what to say, or the interlocutor is tolerant and prepared to help out.

2 Could be expected to perform like this without support in normal circumstances.

3 Could be expected to perform like this even in difficult circumstances, for example when in a surprising situation or when talking to a less co-operative interlocutor.

4 This describes a performance which is *clearly below* his/her level. Could perform better than this.

This rating scale was presented on the front of each questionnaire, and given again in short form at the top of the column for each possible response on each page of the questionnaire, so that it appeared rather like a Likert scale (Oppenheim 1966/92: 195–200).

Subjects

Since the teachers taking part in the study were to be volunteers (with a symbolic reimbursement for their effort), and since no national coordination structure of any kind exists in Switzerland for English Language Teaching, it was obviously impractical to expect to implement a fully representative sampling design. Nevertheless, through contacting teacher associations and exploiting personal contacts a network reasonably representative of levels, sectors and language regions was put together. The political rather than demographic balance was the sought after ideal. In this "magic formula" used to govern Switzerland and decide representation in the *Bundesrat (*collective premiership*)*, the German-speaking cantons get 4 places, the French-speaking 2 and the Italian-speaking 1. Thus the French-speakers have a 28% political representation rather than the 17% representation their demographic proportion would suggest. In the event, the proportion of teachers from the French and German-speaking parts was virtually perfect. Out of exactly 100 teachers who took part in both the questionnaire survey and rating conference, the proportion for the Romandie (French-speaking part) was 28 teachers (i.e. exactly 28%), but teachers in the Italian-speaking region were very under-represented with only 4 teachers (14 would have been politically correct). This was due to the loss of a large group of teachers at the last minute. Two teachers from areas where Raeto-Romansch is the official mother tongue also took part, which with the 67 from the German-speaking cantons makes up exactly 100 teachers. Of these 100, 25 were native speakers.

Several teachers taught in two educational sectors, but overall the proportions were: lower secondary sector (mainly 14–16 year olds) 35 teachers; Berufsschule (mainly 16–18 year old apprentices) 15 teachers; Gymnasium (mainly 16–19 year olds): 19 teachers; Adult education 31 teachers.

The 25 native speakers were distributed fairly evenly around the language regions, but not between the educational sectors. The 35 lower secondary teachers and the 19 Gymnasium teachers were all speakers of the language of their language region. The 15 Berufsschule and 25 Migros Club School teachers were mixed, and the other 6 adult education teachers

(Volkshochschule and University of Lausanne Language Centre) were all native speakers.

Each teacher was asked to select 10 learners, if possible 5 from two different classes. Not all teachers rated 10, however, giving a total of 945 learners in the survey. Teachers were asked to select 5 learners from almost the full range of ability by listing the learners in rank order of English proficiency, excluding the top and bottom learners in the group, taking the second to top and second to bottom learner, identifying a learner approximately in the middle, and then taking two more learners at the mid point between the middle learner and the learner second to top/bottom. This gave a very defined sample and made comparisons between teacher judgements easier as a result. However, there is a possibility that this subject selection design may have reinforced what turned out to be an already powerful tendency to norm-reference and over-discriminate between learners on the basis of the descriptors, which was one of the problems encountered in the data.

The questionnaires were allocated on the basis of information about the number of years and hours a week learners had studied the language, tempered by the experience of the project coordination team. With only 7 questionnaires, and 945 learners, it was possible to arrive at the figure of 100 subjects per form recommended by Madsen (1986: 2) and Jones (1993).

Problems with the Analysis

The main output from a Rasch analysis takes the form of ability/difficult estimates, standard errors and fit statistics. The output of the FACETS programme is discussed in detail at the beginning of Chapter 6 so that it can be more easily related to the scale construction. It very soon became apparent from the rank ordering of ability estimates for learners and difficulty estimates for descriptor items during initial analyses that there were serious problems with the data. Learners with one year of English were coming out higher than learners with 10 years of English; descriptors which were clearly elementary were coming out near the top of the scale, with advanced ones coming near the bottom. In addition, during a one week stay with Mike Linacre at the MESA Laboratory in Chicago at the end of June 1994 with the data from 71 teachers the first analyses run on the data took over 400 iterations to converge using the conventional convergence criteria set as defaults in the FACETS program. An “iteration” is a pass through the data

to adjust estimated values in order to account for the variance. To put 400 iterations into perspective, the default in the program is 100 iterations, and the final successful analyses took only 40–50.

The problems discussed in this section emerged at different points during the process of analysis. For the sake of clarity they are all discussed here, separately from the process of scale construction, which had to be restarted from scratch once corrective measures had been adopted for each of the problems. All the problems encountered are different aspects of one central issue. This study is innovative in that it has combined an itembanking methodology and data collection design usually applied to dichotomous test items, often applying to a limited range of proficiency, to teacher-judged scalar data across the full range of foreign language proficiency. As will become clear in the discussion, an itembanking approach requires a *separate* analysis of forms, whereas a FACETS analysis, in order to adjust for judge severity framework requires an *integrated* analysis so that it can model the measurement framework of the judging situation. Unfortunately it took some time to discover this inherent contradiction in the data collection design, or perhaps in the rather ambitious aim of the study. Some of the problems encountered may have been caused by the data collection design, some appear to have been caused by the way the Rasch model works (or the way Rasch analysis programs work), and some by an apparently unfortunate interaction between aspects of the two.

Use of the Descriptors in the Survey and at the Conference

The first problem was that the initial series of analyses seemed to suggest that teachers reacted to items whilst rating video samples of learners at the rating conference differently to the way they used them to rate their own learners in questionnaires. Separate analysis of the two sets of data showed a *negative correlation* (-0.35) between the calibrations of 32 items common to both the survey and conference questionnaires. The correlation would have been worse had another 8 items not already been removed from the analysis due to obvious extreme "misfit." Misfit is explained in more detail in the discussion of scale construction in Chapter 6. The 5–7 items on the mini-questionnaires were selected to describe the tasks being undertaken (Description and Interaction) and aspects of the quality of the performance (usually Range, Accuracy, Fluency and/or Cooperating Strategies and Pronunciation). What may have happened was that the teachers quickly identified that they were getting the same aspects to rate each time, and used the

0–4 scale to rate on the *aspect* in order to show that one person was better than the other (norm-referencing) despite being asked to rate in relation to the *standard* represented by the wording of the *descriptor*, as in the survey (criterion-referencing). At any rate, an analysis of the conference data alone produced a calibration of the video persons which, though showing signs of a "bonus" for a good elementary performance and a "penalty" for a weak advanced performance, seemed to be a good starting point, whilst the calibration of the descriptors with conference data alone was plain nonsense.

The original intention, discussed with Linacre in 1992, had been to perform a "pooled equating" (Stahl and Lunz 1991) analysis, when all the data relevant to the rating situation is analysed together. This would have given an integrated measurement framework and lower standard errors, but because the conference data appeared to be distorting the survey results, this approach was abandoned in favour of an "anchored equating" design in which the descriptors would be calibrated separately with just the questionnaire survey data. Then the descriptors could be "anchored" at their difficulty values before establishing severity values for the raters and ability values for the learners.

It is also possible that the problem with the descriptors in the rating conference data was caused by the fact that only 5–7 were used for each rating. After the removal of misfitting items on Accuracy and Pronunciation, there were sometimes only 2 or 3 items per questionnaire. As will be discussed in detail later in the chapter, the Rasch model causes a distortion in the estimates away from the middle of the logit scale. Warm (1989) has demonstrated that this problem is particularly serious when only about 10 items are used. The problem is caused by measurement bias, which is "inversely proportional to *n*, the number of the items in the test" (1989: 428) and effectively swamps the true estimation at low levels of *n* when the normal Rasch algorithm (maximum likelihood estimation: MLE) is used. This might go some way towards explaining what happened.

In addition, there were 100 raters at the conference rating 26 learners on 5–7 items with a 0–4 rating scale: 100 x 26 x 5–7 judgements x 4 points = over 40,000 score points; rater severity was being estimated very precisely indeed. However, most questionnaires had 5 items used with at most two learners, a 4 point scale and 100 teachers 5 x 2 x 4 x 100 = a score of only 4,000 points. In retrospect, it is perhaps no wonder that the powerful teacher data (ten times as numerous) swamped the weak questionnaire instruments. Whereas conference data on persons and judges gave sensible

results the mini-questionnaires were rolling each other up when analysed in one integrated run, and hence producing values with a negative correlation to those from the questionnaire survey.

Fortunately, Pollitt (personal communication) considered that one should in any case calibrate the questionnaire survey items separately since (a) in constructing a scale one was interested in how the descriptors were interpreted in practice, not in explanations for stricter or more lenient interpretations and that (b) the basic two facet (item; person) model was a stricter mathematical model which would bring more precise estimates of difficulty. This argument of Pollitt's can be seen as a distinction between constructing a scale and interpreting its use. Linacre puts forward a related point in the FACETS manual (Version 2.8 May 4, 1994: 16–17) in relation to a 4 facet analysis of results to an arithmetic test. The 4 facets are Item, Person, Race and Sex, the former two being measurement facets constructing the scale; the latter two being demographic facets exploring its use to identify possible bias. Two separate analyses are recommended in the example: first to construct the measure, and then to investigate the demographic significance.

In this study a number of demographic facets were included in the data and in the very first analysis run (which took the 400 iterations) all the demographic facets were included. Linacre commented (personal communication) that the effect of having demographic facets active whilst constructing the measure would be for difficulty to be assigned by the algorithm to the demographic factors which should first be assigned to the items (and judges). Pollitt's argument is an extension of this: that by having the judges active during the construction of the measure, one would encourage the algorithm to assign difficulty to the judges which should first be explained through the items. A similar phenomenon is agreed by both Pollitt and Linacre to be the case when using the Rasch partial credit model (PCM: Masters 1982; 1988a 1988b). The PCM allows the algorithm to define an independent rating scale for each item on a questionnaire. One item may encourage just a Yes/No distinction (2 big categories) and not use the other steps on the scale; another item may use all 4 categories equally. The PCM was used in this analysis to investigate whether the 0–4 rating scale was used consistently with the items, but both Linacre and Pollitt recommended against using it to construct the measure since, as a looser model (more things open and moving) it would tend to assign difficulty to the wrong places, i.e. introduce error.

There seems, then, to be a hierarchy of Rasch models from strict to loose in the following order:

- The original two facet (item, person) model used in test item-banking with dichotomous items: most rigorous.
- The two facet rating scale model: with a fixed rating scale e.g. 0–4.
- The many-faceted model: item, person, judge, plus other facets later.
- The partial credit model: all rating scales open, defined by the analysis: least rigorous.

The more rigorous the model, the more precise the calibrations. This argued for putting the conference data to one side, calibrating the descriptors with no account taken for teacher severity, and then returning to the conference data with values for the descriptors from the questionnaire analysis.

Excessive Separation of High and Low Scores in the Analysis

The second problem had been anticipated. It concerned a tendency for the Rasch model to distort calibrations at the top and bottom end of the scale for the test/questionnaire concerned. In a horizontal equating design, this does not matter greatly, and it may well be that a matrix design dampens the occurrence of this phenomenon since the linking is more sophisticated. However, in vertical equating it is a real problem. The effect is that because the ends of the scale for each analysis are distorted, the overlap between all the forms in a vertical equating design are distorted. High scoring elementary students and difficult elementary items are pushed far too high up the common scale created by the overlapping forms, and the reverse happens with low scoring advanced students and easy advanced items. Such distortion, whilst regrettable, is less serious in an itembank because the items will be used in tests in combinations of 30–50 items which will dampen any effects. With a "descriptor bank," however, *any* such distortion is a matter of considerable concern since, in a resulting assessment situation, each descriptor will be used singly.

The problem caused Jones (1993) difficulties during the development of the Eurocentres Itembanker program, which contains a bank of 1,000 items covering a wide range of language proficiency. After discovering the problem, Jones eliminated scores more than 1 Standard Deviation from the

mean (i.e. the top and bottom 16%), but in an experiment at Eurocentres Lee Green in July 1991 to confirm the relationship between the Itembanker logit scale and the Eurocentres scale of language proficiency, information from a bank of C-tests calibrated earlier to the Eurocentres scale (North 1991) plus information from systematic subjective assessments in relation to defined criteria for Oral Interaction and Writing suggested strongly that the distortion started at 20% rather than 16%. Correlations suggested a considerable degree of concurrent validity between Itembanker and the subjective assessments, Writing: Itembanker = 0.90; Oral Assessment :Itembanker: 0.87; n=166; p = < 0.001, so it was decided to restrict the reporting range for the published version of the program to the range between 20% and 80% scores on any test drawn from the item bank.

Jones is not the only person to have come across this problem. Wright and Masters themselves admit (1982: 114) "there can be substantial difference in estimation error between extreme and central scores." Goldstein (1980: 234–5) criticises the Rasch model for the exaggeration it produces in the estimated values for extreme scores stating "the difference in ability for an individual with 99% probability of success and one with a 95% probability of success is about 1.6 units of ability (*logits*), which is the same as the difference in ability between an individual with a 30% probability of success and one with a 70% probability of success. Thus the model discriminates far better at the extremes of the ability scale than in the middle." Camilli (1988: 231) touches on the same point as Jones, but extends it to all IRT models saying that such a ceiling and floor effect make all calibrations outside the central range of -2.0 logits up to +2.0 logits unreliable, though they are truly linear within this range. Talking about the three-parameter Item Response Theory model (estimating guessing & item discrimination as well as difficulty/ability) rather than one parameter Rasch model (estimating only difficulty/ability) Petersen et al (1989) concluded that "measurement error variance for examinees of extreme ability (90% scores) could easily be 10 or even 100 times that for more typical examinees" (Petersen et al 1989: 228). In addition, Choi and Bachman (1992: 66) point out that when Rasch calibrations are compared to those from more complex (two and three parameter) IRT models and item difficulty indices from classical test analysis, it becomes apparent that Rasch underestimates the proportion correct score for difficult items and overestimates the proportion correct for easy items. This might explain the phenomena being discussed since the effect will be to make relatively easy items look even easier, and relatively hard items look

even harder. One can extrapolate that the same will happen with learners: weaker learners will look even weaker, stronger learners even stronger. When several overlapping tests or questionnaires are analysed together, the effect will again be to exaggerate the extent of the overlap, i.e. to push "good" elementary learners and difficult elementary items too far up the scale, and push "weak" advanced learners and easier advanced items too far down the scale.

In the questionnaire data set, the scores on the descriptors themselves were comfortably in the middle range, but scores for learners ranged up to 100%. Warm (1989: 427) comments that in both horizontal and vertical equating "it is assumed that the tests to be linked are administered to examinees for whom the tests are of appropriate difficulty." He proposes a different algorithm (Weighted Likelihood Estimation) and suggests that when using the more usual maximum likelihood estimation (MLE) "rational bounds" should be set to the range of data that is accepted (Warm 1989: 442). If one accepts that Rasch takes as a starting point that tests (here questionnaires) are applied to persons of appropriate ability, then it follows that data from people who appear from their scores *not* to have been of appropriate ability should be removed from the analysis in order to minimalise the distortion. That is to say, only data from test scores in the middle range can be safely retained in the data. The practical question then is: what constitutes the middle range? After experimenting with lower cut-offs, bearing in mind the severity of the problem, a fairly radical interpretation of "rational bounds" was adopted and all learners with scores under 25% or over 75% were removed from the analysis. This meant that descriptors were calibrated on learners whose ability was very close to the difficulty of the descriptors, which increases the accuracy of the calibration (Warm 1979; De Jong personal communication; Pollitt personal communication).

Excessive Discrimination by Teachers (Norm-referencing)

The problem of extreme scores discussed above had been anticipated, but correcting for it still did not remove the problem of a clearly excessive overlap between learners from the different questionnaires. Good elementary learners were still coming out too high, weak advanced learners too low. The fact that this was a real and not an imagined problem was clear from three sources of information:

1. From figures for hours of study. Learners with 80 hours English just cannot be as good as people in a class studying for Cambridge Proficiency.
2. From the placement of the 23 learners for whom there were video performances. Learner who had put in a good lower intermediate performances were calibrated higher than weaker advanced learners. Renate and Rosemarie were coming out at an advanced level, though they would have been unlikely to pass First Certificate.
3. From Itembanker (Rasch-based) test results for 53 adult learners who had also taken Itembanker tests reporting onto the Eurocentres Scale of Language Proficiency. Fortunately, these included Renate and Rosemarie who were placed at Level 5 and Level 5+ on the Eurocentres scale: strong *Threshold* performances, below First Certificate ("C" Pass = approx. Level 6+ (North 1991;1994), and certainly not advanced.

There now seemed to be an interaction between the way the Rating Scale Model (Wright and Masters 1982) works and a tendency for some teachers to use the descriptors to discriminate between learners (norm-reference) rather than to rate in relation to the standard defined in the descriptor (criterion-referencing). The result was an exaggerated range of ability being covered by the judgements of each individual teacher on their class, i.e. the Standard Deviation of the judgements of each teacher was excessive, which meant that the Standard Deviation of the learners on a scale was excessive, which meant that the overlap between learners on the combined scale was still exaggerated. This problem does not seem to be reported in the Rasch literature, though Jones (personal communication) reports having met it in trying to put Cambridge Syndicate subjective assessments onto a common scale, and Pollitt (1994) reports a tendency for teachers in England to spread their class out too far on the 10 level English National Curriculum proficiency scale.

To deal with the problem a procedure based on Standard Deviation (SD) of teacher judgements was adopted. The logit range between the 1st and 5th learners (i.e. the best and weakest) in the samples from each class was calculated. Then the 1st and 5th were taken out of the analysis, and the range of judgement between the three left, (2nd to 4th) i.e. between what the teachers had judged to be the approximate 25% and 75% percentiles, was calculated. The impression gained during an examination of the data which preceded a more detailed analysis had been that a range of 3 logits for

the judgements of each teacher, representing about half the range of logits now covered by a single questionnaire form, seemed to produce reasonable results. Taking the sum of the logit ranges representing the difference between learners 2 and 4, one standard deviation proved to represent 2.978 logits. Remarkably this figure was virtually identical in the second year. Teachers who had a range of logits equal to 2 SDs (i.e. 4.188), of whom there were 5, also had their 2nd and 4th learner removed, their class thus being finally represented in the survey only by one, "typical" learner in the middle of their class. Conversely, for teachers who had a range of logits for all 5 learners less than 2.978 (1 SD on learners 2–4), i.e. for teachers who were clearly criterion-referencing, not norm-referencing, learners 1 and 5 were reinstated.

The effect of all of this on item calibration was as follows. Firstly the space between the top two questionnaires (E and I) was widened by 0.4 logits from 1.1 to 1.5 in an overall combined logit scale of approximately 10 logits. Secondly, the space between the second and third questionnaires from the bottom (W1 and W2) was reduced from 0.66 logits to 0.39. Both these moves appeared to be moves in the right direction. The gap between the top two questionnaires should be wider than the gap between the others since some items representing comprehensive mastery if not near-native competence had been included in the top questionnaire. The second and third questionnaires, by contrast, had been designed in parallel to reflect the *Waystage* level; the difference of difficulty was accounted for by the fact that the one was anchored down to a lower questionnaire, whilst the other was anchored up to a form aimed at *Threshold Level*.

Excessive Overlap in One-Step Equating

After the decision to concentrate on the conference data alone, each of the 7 questionnaires had been analysed separately. The fourth problem related to whether one calibrated the questionnaires separately, and then put them onto a common scale by adjusting for the difference of difficulty between the items on them, or whether one analysed all data simultaneously. Stahl and Lunz (1991) call the one approach "anchored equating" and the other "pooled equating." Glass (1988) calls the former "disjunct scaling" and the latter "multistage testing." Jones (1993: 125) talks of "common-item equating" and "one-step item-banking" whilst Kenyon and Stansfield (1992) prefer the term "concurrent calibration" for the latter. Woods and Baker (1985: 128–9) suggest that one should calibrate each form separately (disjunct

scaling) *by hand*, and Pollitt (personal communication) states that if one compares the two methods with the same data it becomes apparent that Rasch model analysis programs designed for pooled/one-step/concurrent equating tend to exaggerate the true overlap between data collection forms. They ratchet the data from the separate forms too closely together.

Both Jones and Kenyon & Stansfield suggest that it makes little difference which of the two methods is chosen, and that in any case only the calibration of the common "anchor items" is affected. Jones (1993: 152–157) discusses this issue in some detail. He explains that the function of the anchor items common to both forms is to "push" the two forms apart until stable values are reached. He demonstrates that in his dichotomous data the length of the scale produced by each form is fixed early in the analysis process and that the remaining iterations "push" the forms apart to the optimal values. In his example, he shows that the two methods produce the same values for the weaker items on the more advanced form, and the same values for the harder items on the lower form — i.e. *there is no difference to the amount of overlap between two forms.* On the other hand the items at the top of the harder form increase in difficulty value, and those at the bottom of the lower form decrease in difficulty value—i.e. *the overall scale length increases.* He concludes that "one-step" equating is superior in that (a) items are calibrated in a wider framework of reference, taking account of the whole integrated data set from the total measurement situation (which as mentioned above, FACETS is designed to exploit) which leads to lower standard errors and presumably more precise measurement, and (b) the overall scale length—i.e. the degree of separation, the reliability, increases.

However, in this analysis, even after removing the conference data, correcting for extreme score distortion and for excessive norm-referencing as described above, the overlapping effect was still evident. Pollitt's and Jones/Kenyon & Stansfield's opposing views of the matter were checked in relation to this data through contrasting runs sharing exactly the same specifications:

1. Strict "disjunct" equating: the calibrations produced by separate questionnaire analysis, adjusting by hand (in Word tables) for the difference of average difficulty between the questionnaires.
2. "One-step/concurrent/pooled" equating: an integrated analysis with all 7 questionnaires.

3. "Anchored" equating: an integrated run with all 7 questionnaires *but* with items anchored at the calibration values from the separate questionnaire analysis to see what happened to learners.

For this data, Pollitts's view was borne out. In the integrated analysis (2), the overall scale for the descriptors *shortened* from a range of 10.36 logits produced by the separate analysis to give a range of only 8.41 logits, with increased overlap between both learners and items. For example the descriptor *Can express or ask for opinion*, an item placed at elementary level in the Eurocentres content specifications, and in the (University of London-based) European Certificate Project, which had been calibrated in the separate analysis at -0.91 logits, taken to be *Threshold Level* as explained in Chapter 7, was ranked 90th out of 209 items in the separate analysis. In the integrated analysis, it was calibrated at 1.14 logits on the shorter scale, coming in as the 37th item. This would have put it at a considerably more advanced level, which was not credible. In the third analysis, with items anchored at their values derived from the separate analyses, the items of course stayed in the same place, but this did not stop the learners overlapping excessively again. Therefore, disjunct equating was maintained as the methodology for this data.

The contrast with Jones' finding is certainly surprising. Two possible explanations are the following. Firstly, Jones' study on this issue looked at the combination of two tests: where they joined the overlap was the same by both methods, whilst at the ends the scale increased. Yet in a series of 7 questionnaires, the join between the two forms b + c is simultaneously both the join between that pair (b + c) and the join between the preceding and following pairs of tests (a+b) + (c+d). What happens to the increase in scale length beyond Jones' pair of tests? Does it cause overlap with the next pair? Secondly, Jones' study looked at this issue in relation to dichotomous data, while this study employed scalar data. As mentioned earlier, Jones has also discovered hints of excessive rater norm-referencing (spreading learners out too much) in attempts to create a common Cambridge EFL scale. He reports (personal communication) that the Rasch logit scale produced by the Rating Scale Model of judgements seems to be inevitably longer than that produced by the dichotomous model. But the fact that the data in this survey had been "topped and tailed" so severely *without* removing this problem suggests that there might be a problem with the Rating Scale Model itself which perhaps does not occur with the simple dichotomous model. If true, this might be connected with the fact that, in Pollitt's hier-

archy of rigour mentioned above, the RSM is a less rigorous model than the dichotomous model and, since each item has a scale, the possible distortion caused by "extreme scores" may occur with the rating scale *for every item* rather than just for the test score total, as with the dichotomous model.

Linking Data Sets

Whatever the reason for the failure of the "one-step" integrated approach to provide estimated values which could be accepted as plausible for further investigation, the need to keep to "disjunct" equating analysing each questionnaire separately posed problems for linking the data from the Rating Conference with that from the survey so as to be able to take rater severity into account in the estimations of ability for the learners. There were two ways in which this could be done (each with minor variants).

The first way was to analyse the conference data to establish rater severity so as to anchor it for the questionnaire data and then to establish common persons (video people also rated by their class teacher on that questionnaire). Then an integrated analysis of all the data from both data sets with items anchored to the values from the questionnaire analysis, and either teachers anchored to the values from the conference analysis and/or "common persons" (the video people) anchored to their values from the conference analysis. The second way was to conduct separate questionnaire analyses with the survey questionnaire analysed together with the mini-questionnaires containing items from it, with items anchored to their values from the questionnaire analysis and all teachers and learners floating.

In the event the former approach did not work for two reasons. Firstly, in bringing together logit scales produced in two separate analyses, one confronts the problem that the logit scale values produced by a Rasch analysis do not have an absolute value, they are an artefact of the particular analysis and data set which produced them:

> "Since logit measures are estimated from the counts observed in the current experiment, the meaning of a logit in terms of the underlying variable need not be invariant between experiments." (Linacre and Wright 1989: 4)

The problem was that a point which in the circumstances can only be described as "optimum teacher misfit" was quickly reached after only a few analyses. The removal of misfitting items or persons, or judges, causes other items or persons or judges to start to misfit more in relation to the more

clearly delineated dimension. The clearer delineation also leads to an increase in scale length. Whilst such increased scale length, entailing increased separation and so higher reliability, is to be welcomed in a main analysis, in a *supplementary* analysis, being undertaken as an adjunct to a main analysis it caused a real problem as the top and bottom learners moved past the top and bottom items (which were anchored) on the scale. Removing teacher misfit of 1.8 or more led to other teacher to misfit to the 1.8 level and caused the top learner left in the analysis (Sibylle) to climb up the supposedly common logit scale past the anchored items to an extent that just was not credible. Clearly the data was behaving as a separate experiment in terms of Wright's statement cited above. The solution adopted was to set the number of iterations at 100, the default in the program.

However, even when this problem had been solved, the integrated "anchored" analysis of the questionnaire and conference data together exaggerated the overlap between forms and mixed the learners up excessively as usual. An attempt to force the issue by also anchoring the learners in the 10 video performances with the best fit statistics as well as the items and teachers produced massive initial misfit on the facet "Occasion" (Survey/Conference) and a failure to converge. Therefore solution (b) disjunct equating was adopted.

Although the main aim of this study was to calibrate the descriptors for the Common Framework and the Language Portfolio, it was felt necessary to bring the two data sets together and obtain ability estimates for the learners to check two things:

Firstly, how the items were actually *used* to rate learners. A scale of just the descriptors could be organised into levels, but unless one could match learner performances up to those words one would not really know what the levels were *actually* describing. The ability estimates of the learners clarified what had been a point of concern about possible shrinking in the scale and confirmed that the scale was in fact more linear, more equal interval, than had first been thought.

Secondly, whether data from two contrasting rating situations would produce similar ability estimates and placement at the same level on a proficiency scale for those learners common to both rating situations. The two situations were: (a) *Questionnaire*: continuous assessment based on a class teacher's perception of performance at the end of a school year, and (b) *Rating conference*: a summative assessment based upon a performance in a 5–15 minute video recording. If descriptors *do* seem to be interpreted differ-

ently in different contexts, it would at least be interesting to know it, especially in relation to a common framework. This issue is also discussed in talking about estimates for learner achievement in Chapter 8.

Summary on Analysis Problems

Jones (1993) comments that calibrating items with the Rasch model is more of an art than a science. Linacre (1990: 7) cautions that deciding whether data fits the measurement model is a matter of informed judgement, not of reading numbers from a computer print out. Experience with this study bears out both views: Rasch is no panacea, there is no simple answer delivered effortlessly and "objectively" by the analysis procedure. Judgement at all key points in the process is required. Fortunately, in working with descriptors from known sources which have been through a pre-testing process, and working with learners on whom one has some other information, in a subject area one is familiar with, one at least has transparent material from which to make a considered judgement. It is probably a fallacy to think that any empirical analysis can be "objective" in the sense of being totally independent of judgement. Despite the difficulties encountered, it has been apparent that it is a feature and a strength of the Rasch model that by working at an item level, one can, through judgement, refine down the data collected in order to identify and exclude things about which, on reflection, one can say nothing, and to adjust for the imperfections inherent in any data and the distortions probable in any analysis. The problems with scale distortion (non-linearity) discussed above are not confined to the Rasch model, and adjustments can be made for them. It is also true that the Rasch model is very robust (Forsyth et al 1981: 185), that its limitations actually lead it to tend to throw out as misfitting more than it ought to (Choi and Bachman 1992: 74) and that Rasch model itembanks which identify and exclude data to minimise distortion (for example Itembanker) can produce reasonable correlations to both systematic subjective assessment and public examinations (North 1994).

It is however, certainly curious that with the exception of the reservations about scale distortion voiced by Goldstein (1980), Camilli (1988), Petersen et al (1989), Warm (1989) and Choi and Bachman (1992), the kind of problems which occur with vertical equating discussed here do not seem to figure much in Rasch or general IRT literature, which tends to concentrate warnings on selecting the appropriate IRT model and on unidimensionality. Neither Madsen (1986) in his discussion of the development of a

ESL Rasch itembank and computer-adaptive test, nor Stansfield and Kenyon (1992) in their discussion of vertically equating tests for Chinese mention these problems. Madsen (1986: 15) does state that it is necessary to do disjunct equating, like Woods and Baker (1985: 128), whereas Kenyon and Stansfield come to the opposite conclusion. Engelhard and Osberg (1983), in their discussion of vertical equating in relation to reading tests, give no hint of the problems encountered by Jones (1993) or in this study stating (1983: 291) that the only criteria which need to be satisfied for Rasch vertical equating are that the items fit the model satisfactorily, that the anchor items are linearly related (i.e. pushing in the same direction), and that the network of forms must be adequately and consistently linked throughout. Finally Lee (1992: 203) in discussing the vertical equation of reading and maths tests through initial separate analysis of tests to remove misfit followed by one-step equating of vertically linked forms states that what he describes as "range restriction problems sometimes reported in the literature" can be overcome by a large number of iterations to meet a tight convergence criterion (0.5 score points).

As will be seen in the discussion of the analysis steps in the next section, it can be claimed that the scale of descriptor items meet the constraints listed by Engelhard and Osberg, and the convergence criterion mentioned by Lee, yet the problems still occurred. Linacre states (Personal communication) that it is impossible for messy, real world data to ever fit a mathematical model. As he comments in the FACETS manual on the question of global data to model fit).

> "Neither global fit nor global misfit can be decisive for accepting or rejecting the data. Empirical data is always good in part and bad in part. The useful question is: "Which parts of these data fit well enough to be useful, and which do not?" or "Can these data be edited to bring them to useful form?" The INFIT and OUTFIT statistics *(for each element of each facet)* address these questions." (Version 2.8 May 4 1994: 61)

The strength of the Rasch approach is that by identifying and excluding some of the messiness one can arrive at a generalisable result which is scale-free, test-free and (provided groups can be considered sub-sets of the same statistical population) sample-free measurement. The process by which that has been achieved in this study is outlined in the next chapter.

6 Constructing the Scale

As indicated earlier, the analysis has been undertaken with the many-faceted Rasch Rating Scale Model (Linacre 1989). In a simple Rasch analysis of dichotomous (right/wrong) test items, there are two facets: "person" and "item." When the severity of the rater is taken into account in the calibration, as is possible with the FACETS program, "judge" becomes a third facet. It is also possible to design other facets in the measurement situation such as "occasion" in order to, for example, investigate the stability of the severity of examination judges of at different administration sessions in the same examination (Lunz and Stahl 1990; Stahl, Lunz and Wright 1991; Stahl and Lunz 1992). In this case the two occasions are (1) the rating by each teacher of 10 learners from their own classes and (2) the rating by each teacher of video performances of 23 of the learners in the survey at the rating conference.

It is also possible to include demographic facets in the data so that once the measurement framework has been established, and values attributed to the measurement facets (here: Teacher, Occasion, Learner, Item), the effects of demographic information like age, sex, race (here: Educational Sector, Language Region) can be calculated. As well as overall measures for the demographic facets defined, it is also possible to run Bias studies in order to identify bias against specific individuals or bias by certain items (here descriptors) for or against specific groups, which would suggest treating them with caution in a common framework scale. Finally, since it is very easy indeed to remove a facet from the analysis by varying the "model statement" there is an advantage to defining as a facet any variable one may wish to manipulate during the analysis.

Analysis Specifications and Data Organisation

Given these general possibilities, the present data set was organised in 10 facets as listed below. The 4 measurement facets are marked in italic, with

the demographic facets in normal print. The abbreviations used in Table 6.1 are on the right.

1. Teacher	*Judge*	T
2. Teacher Educational Sector		T Edsec
3. Teacher Mother Tongue		T MT
4. Teacher Language Region		T LR
5. Occasion	*Occasion*	Occ
6. Learner	*Person*	L
7. Learner Educational Sector		L Edsec
8. Learner Mother Tongue		L MT
9. Learner Language Region		L LR
10. Item	*Item*	Item

Data is coded in lines, each line including all facets, for example as shown in Table 6.1.

Table 6.1: The Structure of FACETS Data

T	T Edsec	T MT	T LR	Occ	L	L Edsec	L MT	L LR	Item	Scores

19	1	1	1	1	7	1	1	1	1–25	0,1,2,2…

The program accepts ASCII data expressed as comma-separated-values, which can be produced by converting Word or Excel tables to text files. The same line of data then looks like this:

19,1,1,1, 1, 7, 1, 1,1, 1–25,0,1,2,2,0,2,2,1,2,2,2,2,2,3,2,1,1,2,2,2,0,2,1,2,2

The specification file lists the elements in each facet (each learner is an element in the facet "Learner"). For example, the facet Sector is defined as:

2,Teacher's Sector,A
1=Lower Secondary,0
2=Upper Secondary,0
3=Berufsschule etc.0
4=Adult,0

The four elements are numbered, and because "Sector" is a demographic facet not wanted for scale construction, it is anchored (the "A" after the comma) with each element given the value of zero, in order to make it inactive. In the specification file, elements in a facet can also be grouped. For example the items (here descriptors) can be grouped by content strand. Grouping can be used to print out results separately, and. if desired, to play around with the weighting of anchoring values. Otherwise the specifications determine various options set as defaults, the appearance of the data in the output file and, if desired, score files produced.

The program never builds a matrix, but matches the measurement model(s) specified in the specification file to each line of data in the data file(s) in linear fashion.

The main model statement used in the analysis was:

?,X,X,X,1,?,X,X,X,?,Normal,, ;School Questionnaires with normal scale

which means:

Find a rating by any *Teacher*—ignoring for the moment the demographic facets 2, 3 & 4 (marked with Xs)—on *Occasion* No 1 (School Questionnaires) about any *Learner*—ignoring for the moment the demographic facets 7, 8 & 9 (marked with Xs)—on any *Item*, and use the *Rating Scale* defined as "Normal."

In fact, since the teachers were ignored by being anchored to 0.0 severity in order to calibrate the items, the actual model used for much of the analysis was:

```
X,X,X,X,1,?,X,X,X,?,Normal,,  ;School Questionnaires normal scale
```

with only the two basic facets active: Learner and Item, simulating a two facet Rating Scale Model analysis program.

It is a strength of FACETS that model statements can be written allowing one to define several different rating scales for groups of items, to allow partial credit scoring (each item defining its own scale depending how it is used).

FACETS Output

FACETS produces a wealth of very complex diagrammatic output, and only a basic overview of the tabular information containing values will be given here. This comes in two forms: an All Facet Overview and the Facet Report. The latter is a detailed report on each facet listing all elements in each facet with their scores and statistical properties in the way defined by the user in the specification. An example of an All Facet Vertical Overview is shown in Figures 6.1 for the conference data, chosen because it shows the three principal facets Teacher, Item, Learner.

Figure 6.1 FACETS All Facet Vertical Summary

```
-------------------------------------------------------------------
|Msr|-Teacher                  |+Learner |-Items             |S.1|
-------------------------------------------------------------------
+ 4 +                          +         +                  +(4)+
|   |                          |         |                  |   |
|   |                          |         | Fluently-effortles| 3 |
|   |                          |         |                  |   |
|   |                          |         | Clear well-structu|   |
+ 3 +                          + G-SIBYLL+ Descr: clear detai+   +
|   |                          |         | Comm spontaneous  |   |
|   |WillisA                   |         | Clear natural pron|   |
|   |                          |         | Only occasional sl|   |
:   :                          :         : Relate skilfully t:   :
|   |Muller                    | M-DORIS |                  |   |
+ 2 +                          +         +                  +   +
|   |                          |         |                  |   |
|   |Potts     Von Arx         |         | Adjust to changes |   |
:   :                          :         : Natural turntaking:   :
:   :                          :         : Stretches of langu:   :
|   |Margot                    |         | Help discuss along|---|
|   |Aebischer Hefti     Honegg |         |                  |   |
:   :Kaufmann  Pollak          :         :                  :   :
+ 1 +Gantenbei Richon    Sutter +         +                  +   +
|   |Batliner  De Souza  Gyger  |         | Account for views |   |
:   :Humphry   Kretz     Maier  :         :                  :   :
:   :Miconet   Urech           :         :                  :   :
|   |Arman     Boni      Buhaye | G-VIRGIN|                  |   |
:   :Gerber    Hadley    Jungen :         :                  :   :
:   :Messerli  Scheurer  Schmid :         :                  :   :
:   :Schuepp                   :         :                  :   :
|   |Braun     Chadburn  Klingen|         |                  |   |
:   :Koppel    Taylor           :         :                  :   :
|   |Durr      Gillespie Ries   | G-CHRIST| Pron clearly intel|   |
:   :Ritter                     :         :                  :   :
* 0 *Bailly    Lendenman North  *         * No misunderstandin*   *
```

Development

of

a

Common

Framework

Scale

of

Language

Proficiency

:	:Schlapbach	Schnabl	Turin	:	:	: :
|	|Bell	Cavelti	Felix	| M-RENATE	| Reasonable accy	| 2 |
:	:Geiser	Heinz	Joss	:	:	: :
:	:Mackenzie	Marti	Orsi	:	:	: :
:	:Sager	Tang	Willis	:	:	: :
|	|Andrey	Chuffart	Dolci	| G-MARINA	| Keep going compreh	| |
:	:Hale	Hermle	Meili	:	:	: :
|	|Eberle	Monney	Raas	| M-ROSEMA	| Detailed accounts	| |
:	:Schwager	Stahel		:	: Discuss topics of	: :
:	:			:	: Link into connecte	: :
:	:			:	: Relate plot of boo	: :

Figure 6.1 (cont.) FACETS All Facet Vertical Summary

```
-------------------------------------------------------------------
|Msr|-Teacher                   |+Learner |-Items             |S.1|
-------------------------------------------------------------------
|   |Hug       McDonald  Ommerl | V-THERES| Wide range sim lan|   |
:   :Silva                     :         :                   :   :
+-1 +Grossmann Munro     Oberhoe+ B-MARLEN+ Reas acc- Repertoi+   +
:   :Senn      Zeller          :         :                   :   :
|   |Burton    Fluckiger Hubsch |         | Turn:  st/mai/end |   |
|   |Cormon    Corry     Pequin |         | Descr: - events/ac|   |
|   |                           |         | Descr pets/possess|   |
|   |Regan     Straessler       |         | Explain (dis)likes|---|
:   :                           :         : Repertoire of basi:   :
+-2 +                           + M-MARCEL+ Con: initiate/ mai+   +
:   :                           :         : Descr: extended   :   :
:   :                           :         : Desr: acts & exper:   :
|   |                           |         |                   |   |
|   |                           | V-FLOREN|                   |   |
|   |                           |         | Descr selves &    |   |
|   |                           |         | Limited repertoire|   |
+-3 +                           +         +                   +   +
|   |                           |         |                   |   |
|   |Robert                     | B-PASCAL| Ask/answ: Selves  |   |
|   |                           |         | Descr where live  | 1 |
|   |Glanzmann                  | S-NICOLE|                   |   |
+-4 +                           +         + Interact simply   +   +
|   |                           |         |                   |   |
|   |                           |         |                   |   |
|   |                           | S-LORENZ|                   |   |
+-5 +                           +         +                   +(0)+
-------------------------------------------------------------------
```

In this analysis, the items were anchored at their values in the questionnaire survey, but that does not appear on this overview. On the left one sees the logit scale, running from -5 up to +4, and on the right one sees an adjusted version of the rating scale used. Since the rating scale is defined by performance conditions, the learners, in the middle in capitals, with an initial noting their educational sector, are placed alongside the descriptors which describe what they *could be expected to do in normal circumstances*. Thus Marina is said to be exactly at the level when under normal circumstances she could: *keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production*, but one would not really say that she could: *communicate with reasonable accuracy in*

familiar contexts; generally good control though with noticeable mother tongue influences like Renate can, though she could certainly *give detailed accounts of experiences* and *can discuss topics of interest.*

The All Facet Vertical Summary is thus a very useful tool for (a) seeing if what is coming out makes sense, if you have some means of knowing how good the learners actually are, and (b) for giving an overview of a final result, as now. On the left one finds the teachers, of whom Bailly, Lendeman and North (my wife, not me) come out with perfect severity i.e. 0.0. WillisA and Muller, on the other hand come out as amazingly strict, and Robert and Glanzmann come out as extremely lenient. Most of the other teachers land in the range between -1 logit and +1 logit, so the bulk of the teachers are covering a range of 3 logits. However, altogether the teachers cover a range of 7 logits compared to the range of about 8 logits covered by these items and 8 logits covered by these learners. Hence the interest in taking account of judge severity in making statements about the range of achievement of different educational sectors based on a questionnaire survey.

The main FACETS reporting mechanism, however, is the Measurement Report given for each active facet. Figure 6.2 shows this report for the 14 learners in the Conference All Facet Vertical Overview. By this stage, 5 of the original 23 learners had been removed because they had obtained "ex-

treme scores" over 75% or under 25% on the mini-questionnaire concerned, and there are therefore no Observed Average scores on the 0–4 rating scale (third column above) greater than 3 or less than 1. The logit calibration, which determined the learner's position on the All Facet Vertical Overview is in the fifth column. The model error in this calibration, also expressed in logits, is quite small here because each calibration is based on about 5 judgements by 85 judges. For the descriptor values calibrated from the survey questionnaires, the model error is higher at around 0.25.

The next four columns give the fit statistics: the difference in the estimated calibration between what the model expects (if life were perfect) and what happens in the data. There are two main statistics: (i) the *mean square* being the *amount* of misfit, and (ii) the *standardised residual*, a measure of *plausibility* testing the statistical hypothesis: "Do these data fit the Rasch Model exactly? Since no empirical data ever does, results must be interpreted with this in mind" (Linacre: FACETS Manual Version 2.8: 64). The *mean square* statistic expects a result of 1: more than 1 is misfit, less than 1 is "overfit:" a result a bit too good to be true. Different writers cite cut-offs of differing severity which they apply to define misfit, but the range 0.5 to 1.5 is conventionally considered "okay." The *standardised residual* is expressed as a standardised score, i.e. with zero in the middle (as for the logit scale). Anything over 2 on a standardised score is under the conventional statistical criterion of 95% significance: it would be expected to happen only 5% or less of the time, so it seems somewhat implausible. During the process of reducing misfit the *mean square* (*amount*) is probably more useful (Linacre personal communication). In determining the unidimensionality and hence model fit of the final data set and values, the *standardised residual* (*plausibility*) is more useful (Hambleton et al: 1991: 66; Stansfield and Kenyon 1992: 10).

A mean square and standardised residual is offered for two fit statistics: for INFIT and for OUTFIT. The INFIT statistic is sensitive to unusual patterns around the range of ability of the learner, whilst the OUTFIT statistic is the figure reported in all Rasch analysis programs giving the extent to which the learner "got wrong" items whose value is estimated to be below his/her level of ability and the extent to which he/she "got right" items

supposed to be above his/her level of ability, though here we are talking about ratings.

Figure 6.2: FACETS Learner Measurement Report

```
Swiss Language Framework Survey:
Table 7.6.1 Learner Measurement Report (arranged by 6mN)
-----------------------------------------------------------------------
Obsd Obsd Obsd  Fair | Calib Model | Infit    Outfit |
Score Count Av  Av   | Logit Error |MnSq Std MnSq Std|Num Learner
-----------------------------------------------------------------------
973   380   2.6  2.9 |  3.05  0.10 | 1.4  4  1.3  3  | 547 SIBYLLE
697   351   2.0  2.7 |  2.19  0.10 | 1.4  5  1.5  5  | 533 DORIS
650   304   2.1  2.2 |  0.55  0.10 | 1.2  2  1.2  2  | 650 VIRGINIE
614   299   2.1  2.1 |  0.28  0.11 | 1.3  3  1.0  3  | 651 CHRISTIA
668   271   2.5  2.0 | -0.17  0.11 | 1.5  5  1.5  4  | 196 RENATE
334   148   2.3  1.9 | -0.36  0.15 | 1.5  3  0.5  3  | 639 MARINA
639   272   2.3  1.9 | -0.57  0.11 | 1.4  4  1.4  4  | 197 ROSEMARI
543   243   2.2  1.8 | -0.80  0.12 | 1.2  2  1.2  2  | 606 THERESE
1163  556   2.1  1.8 | -0.92  0.08 | 1.2  3  1.2  3  | 631 MARLENE
449   224   2.0  1.4 | -2.01  0.12 | 1.4  3  1.4  3  |  22 MARCEL
410   238   1.7  1.3 | -2.49  0.12 | 1.1  1  1.1  1  | 605 FLORENCE
737   546   1.3  1.1 | -3.40  0.08 | 1.1  1  1.1  1  | 630 PASCAL
690   387   1.8  1.0 | -3.70  0.09 | 1.1  1  1.1  1  |  66 NICOLE
574   388   1.5  0.8 | -4.71  0.09 | 1.1  1  1.1  1  |  65 LORENZA
-----------------------------------------------------------------------
652.9 329.1 2.0  1.8 | -0.93  0.11 | 1.3 3.2 1.3  3.2|Mean Count 14
206.3 112.2 0.3  0.6 |  2.10  0.02 | 0.1 1.3 0.1  1.3| S.D.
-----------------------------------------------------------------------
```

```
RMSE  0.11  Adj S.D.  2.10  Separation 19.53  Reliability 1.00
Fixed (all same) chi-square: 6645.3  d.f.: 13  significance: .00
Random (normal) chi-square: 13.0  d.f.: 12  significance: .37
```

By this stage, 4 of the original 23 video people had been removed because of large amounts of misfit, so the *amount* of misfit in this data set is not too worrying. However the standardised residuals are high (above 2) for half the sample indicating a relatively poor model fit. In this case, this could well be a reflection of the fact that the items were anchored to values established on another Occasion (survey questionnaire) in a different form of assessment, and represents one of the problems in linking the two sets of data. As mentioned above, Linacre himself prefers a qualitative approach to analysis rather than the setting of arbitrary statistical criteria. Misfit/fit is a continuum, data never fits a mathematical model, life is messy. Some kinds of data will have more misfit than others and the aim is to reduce the amount of mess and achieve clarity through a rolling series of analyses to produce a result which can be judged acceptable or unacceptable. As discussed in Chapter 5 the conference data was put to one side in order to investigate the performance of items in the survey questionnaires through such a process as discussed in the rest of this chapter.

Investigating Use of the Rating Scale

One of the first points to check in a questionnaire analysis is that the rating scale attached to each item is actually used in a sufficiently similar way to justify regarding it as one single scale. It might be, for example, that some items encourage a Yes/No distinction, or some other distinct pattern of scale use. It might be that some teachers use the scale in a personal way—never using the middle category, for example. Like most Rasch scalar analysis programs, FACETS allows analysis of the way the rating scale is used. It does not give nice diagrams like some programs, but it does provide information in bar chart form which makes it easy to "eyeball" and compare the patterns used.

Two series of partial credit analyses were run very early in the analysis. The first was a partial credit analysis allowing each teacher to define their own rating scale by the way that they used it, therefore giving 100 variants of the rating scale. The second was a partial credit analysis for each survey questionnaire allowing the scale for each item to be defined through the process of the way the particular item was used, rather than assuming that the pattern of responses would be the same for each item.

Teacher Scale Use.

This did not show any noticeable differences of usage apart from the fact that a small group of 4 teachers tended not to use the middle step (2: Normal circumstances), preferring to make definite choices between Step 1 (In favourable circumstances) and Step 3 (Even in difficult circumstances). In the FACETS output bar charts the difference showed up as follows. Normal use of the rating scale looked like the first example in Figure 6.3, chosen at random and actually representing the use of the scale by Teacher No 11.

Three representations are given. According to the FACETS Manual (Version 2.8. May 94: 68):

- The *Mode* presents the most probable category, the number of the scale step is printed where it starts to be the most probable category.
- The *Median* (also called a Thurstone Threshold) represents the point (the threshold) at which the scale step starts to be used; therefore the step number appears at the far left of its "zone of use" i.e. further left than for the Mode.
- The *Mean* given the expected logit score for that scale step.

Figures 6.3 shows astonishingly regular use of the scale.

Figure 6.3: Typical Use of the Rating Scale

```
Scale structure FOR "CONDITIONS" Teacher 11
Relative
Logit:-5.0  -4.0  -3.0  -2.0  -1.0  0.0   1.0   2.0  3.0  4.0   5.0
        |     |     |     |     |    |     |     |     |    |    |
  Mode:<0------01------------12------------23----------34--------4>
Median:<0------01------------12------------23----------34--------4>
  Mean:<--0----+-------1------+-----2------+------3------+----4--->
        |     |     |     |     |    |     |     |    |    |     |
Logit:-5.0  -4.0  -3.0  -2.0  -1.0  0.0   1.0   2.0  3.0  4.0   5.0
```

There was little departure from this pattern except for 4 teachers. For one of these four teachers, the same output is given in Figure 6.4. As the Mode shows quite dramatically this teacher is not *likely* to use the scale Step 2. The most probable scale step is nearly always something other than 2. The Median shows that the zone of use of Step 2 is considerably smaller than for the other steps. Basically, this teacher tends to use steps 1 and 3 resorting to 0 and 4 when necessary, but rarely using a 2. Yet, as the regularity of the Mean shows, this has very little effect on the outcome of the teacher's judgements, and probably has no effect on either the calibration of the learner or the estimate for the teacher's severity.

Two of the four teachers showed INFIT and OUTFIT of Mean Square 1.3. (0 std), one shows 1.5 (2 std) for both, and this particular example, Teacher No 41 shows INFIT of 1.8 (3 Std) and OUTFIT of 1.9 (3 Std).

The fact that all the 4 teachers showed “noisiness” or “misfit” when the standard rating scale was used indicated that it might be an advantage to allow them to use their own version of the scale before coming to a final decision on teacher severity.

Figure 6.4: Unusual Use of the Rating Scale

```
Scale structure FOR "CONDITIONS" Teacher 41
Relative
Logit:-5.0  -4.0  -3.0 -2.0  -1.0   0.0   1.0  2.0   3.0   4.0  5.0
        |     |     |     |    |     |     |    |     |     |    |
  Mode:<0------01-----------------12-23------------------34-----4>
Median:<0------01---------------12-----23----------------34-----4>
  Mean:<-0-----+---------1------+----2----+------3---------+-----4>
        |     |     |    |     |     |     |    |     |     |    |
Logit:-5.0  -4.0  -3.0 -2.0  -1.0   0.0   1.0  2.0   3.0   4.0  5.0
```

Item Scale Use

The same operation was also repeated for all 280 items, again comparing the bar charts and grouping the items according to shape. After this initial analysis, what appeared to be 10 basic patterns were identified, given names, and then given their own rating scales. This is done by writing a model

statement in the specifications. The 10 exploratory rating scales for items were as follows:

1.	LoadOne	a difficult item: peak on Step 1
2.	LoadTwo	the most frequently used pattern: normal
3.	LoadThree	an easy item: peak on Step 3
4.	Something23	a slightly less marked version of LoadThree
5.	Zero23	little use of 0 or 1: most responses 2 or 3
6.	ZeroTwo	either they can do this or they can't
7.	ZeroThree	as above: but stronger
8.	Bridge	little use of 0 or 4
9.	Camel	two peaks: little use of Step 2
10.	Flat	everything gets used

The results were extremely low key. There was, for example, no noticeable improvement in the calibrations and no apparent effect on model fit. The fact that the so many variants of scale use were present was normal in a rating situation. There was no suggestion that items of a particular type tended to be rated in a particular way. Accordingly, Linacre (personal communication) suggested that the loss of precision by importing error which would result in use of the partial credit model would more than outweigh any negligible increase in precision by grouping items under scale variants. The conclusion reached was that the use of the rating scale was sufficiently general to justify keeping the one original rating scale for all items.

Dimensionality: Identifying Problematic Content Strands

An essential part of the process of identifying and excluding misfit concerns the exclusion of items, and groups of items, which, on reflection, should not be in the analysis at all because they are about something else. This is the issue of "unidimensionality." Any test or assessment which reports a single result acts as if the test items are about more or less the same thing. In classical test theory, a test assessing lots of different things will show low reliability, because it does not separate people out into a rank order it can be

relied upon to repeat next time. Rasch fit statistics, especially standardised residuals according to Hambleton et al (1991) and Stansfield & Kenyon (1992), and a separability statistic, which is a version of a classical reliability coefficient, are indications of unidimensionality.

As discussed in Chapter 3 unidimensionality is a relative concept. A construct which demonstrates a satisfactory degree of psychometric unidimensionality may very well be made up of separately identifiable content areas or "strands" of the dimension. Some of these strands will appear from fit statistics to be more central to the construct being measured, others will show themselves to be less so. An important part of a Rasch analysis consists of "constructing the construct" (Wright: personal communication): honing down the content to a construct which could be regarded as sufficiently (psychometrically) unidimensional for analysis purposes. In this study, the focus was on spoken interaction, so the vast majority of the descriptors defined stages of attainment in different aspects of spoken interaction. Listening was included only in relation to listening during interaction and reading was excluded altogether. However, as well as the descriptors on spoken interaction, descriptors were included on spoken *production* (sustained monologues), some described *written* interaction (correspondence) and yet others described *written production* (writing reports and essays). In addition, while most of the descriptors had been selected or edited to be

applicable to a broad range of general contexts others described proficiency in activities more associated with adult professional life (e.g. meetings, presentations, formal correspondence). Finally, descriptors were included not just for communicative activities themselves, but also for aspects of quality in spoken performance (pragmatic and linguistic), as well as for aspects of socio-cultural competence and strategic competence.

The reason for the inclusion of descriptors which could be expected to be less central to a construct focused on spoken interaction was the project aim to provide a reasonably comprehensive bank of calibrated descriptors for a profiling grid. The hope was that, as in other language testing studies (e.g. Henning 1985), foreign language proficiency would prove to be a sufficiently robust construct to cope with the obvious psychological multi-dimensionality implied by the above and display enough psychometric unidimensionality to justify calibration of descriptors in a large number of the content strands concerned.

However, the inclusion of less central content strands posed a number of questions:

- Would listening-in-interaction fit as part of a spoken interaction construct, or would it prove to behave differently?
- Would spoken *production* fit with spoken *interaction* as part of a speaking construct, or would it behave more like writing?
- Would *written* interaction fit with spoken interaction as part of an interaction construct, or would it behave more like written production?
- Would written production fit at all?
- Would specific purpose descriptors fit with general purpose ones?

Clearly Problematic Content Strands

Six of the original content strands were identified very quickly as being clearly problematic in that a significant proportion of the items classified in the strand concerned showed very large amounts of statistical misfit. These

6 areas, which had to be removed from the analysis as they did not fit in the construct, were the following:

1. Socio-cultural Competence
2. Telephoning
3. Meetings
4. Formal Presentations
5. Independence (a) Need for Interlocutor Adjustment, (b) Need to get Clarification, and (c) Need for Help
6. Pronunciation

Socio-cultural Competence is very probably a genuinely separate dimension. That 10 of the 12 socio-cultural items should have misfitted very substantially was no great surprise since (a) it had been found difficult to formulate descriptors in the first place, (b) it was the area which had proved most problematic in the workshops with teachers, and (c) other studies (e.g. Bachman & Palmer 1981; Pollitt and Hutchinson 1987) report that socio-cultural / socio-linguistic competence appears to form a separate dimension to language proficiency.

Telephoning, Meetings and Formal Presentations were peripheral content strands which have two factors in common. Firstly they are areas

associated predominantly with the world of work and only three or four of the classes in the survey were following courses which could be described as work-oriented, since the language courses for apprentices in the vocational education sector were discovered to be exclusively general purpose language learning. Secondly, these three content strands are areas which teachers have little opportunity to observe behaviour in normal classrooms. Teachers were in effect being asked to judge—or rather guess—about areas of their learners' performance which were outside their experience. Following Smith & Kendall's (1963) concept of behavioural expectations, which had been incorporated in the wording of the rating scale (*Could be expected to perform like this in* x *circumstances*), it had been hoped that teachers would be able to generalise from the evidence they did have about probable ability in areas which they had not actually observed. Whilst this ability to generalise seems from the evidence provided by fit statistics to have worked in relation to, for example, Service Encounters (getting information, using facilities, shopping etc.) where a link to simulation scenarios in course material could be expected, teachers' ability to *consistently* generalise about telephoning, meetings and presentations was shown to be inadequate. Griffin (1989, 1990a) reports a similar problem in relation to his primary reading scale with items connected to library skills, which were also areas of expertise outside his teacher/raters' experience and which he also had to exclude from his analysis. In this case, the removal caused a loss of a further 12 items (Telephoning: 5; Meetings 4; Presentations 3).

This issue is not an "all or nothing" affair and will be touched upon again in discussing suspicions about other content strands later in the chapter. Three descriptors for Service Encounters, for example, did show "noisiness" (marked but not excessive misfit) and two of those three descriptors concerned the use of public transport, something teachers would also have been unable to observe. Three of the eight descriptors for Interviewing also showed "noisiness" and therefore Interviewing was looked at rather carefully. Statistical information (here misfit) can guide decisions about which strands to include and exclude, but a judgement whether to exclude a whole strand or just the worst items in it has to be made sub-

jectively. For these three areas the issue appeared clear cut. For others like Interviewing and Formal Correspondence, both of which can be seen as rather adult and possibly work-oriented activities, suspicions were aroused, but the strands were not rejected at this stage of the analysis.

Independence was from the beginning a rather problematic, though interesting, category. Actually it is about dependency, need for accommodation: simplification and reformulation from the interlocutor (Need for interlocutor adjustment, Need to get clarification) and need for support when trying to say something (Need for help). Despite the choice of a positive title, the concept of "Need for……" is a negative one: more of it is bad, negatively correlated to proficiency and operating in the opposite direction to categories like "fluency" or "accuracy" for which more is good. Most of the Independence items started "Native speakers need to……" or "Needs to……" which sits uneasily with a rating scale in which 2 is "Normal circumstances" and 3 is "Even in difficult circumstances." One or two teachers complained about precisely this incompatibility.

Negative concepts do not necessarily have to have a negative wording (I cannot do x). Heilenman (1990) discusses Bachman & Palmer's (1989) suggestion of descriptors starting "I have trouble……" as an example of a lexically rather than structurally negative formulation. Heilenman (1990:

176) cites scaling research (Schmitt and Stults 1985) which states that when positive and negative formulations are mixed in the same questionnaire, some respondents will ignore the reversed direction of the rating and thus contaminate the data. A factor identifiable as being accounted for by negatively worded items will occur when as few as 10% of the respondents fail to notice the reversal. This issue is also touched upon in the FACETS manual with the suggestion that data may need to be recoded to remove the problem—though that raises the problem of how you know which data to recode! The clash between the implicit negativity of these items and the rating scale produces sort of double-negative, and the avoidance of double negatives is classic advice on questionnaire wording (e.g. Oppenheim 1966/92: 128). It confuses some people and it is difficult if not impossible to find out who they were.

Investigation of the questionnaire papers themselves confirmed the suspicion that some teachers were reversing the direction of their rating, that others were not, and that with others it was hard to tell. The result was inconsistency, which shows up as substantial misfit, as in the two examples in Table 6.2.

The concept involved was not lost entirely, however. Some items, which showed no particular misfit, incorporated a statement about independence as a proviso attached to the end of a *Can do* statement. All of the lower level descriptors on Comprehension in Interaction contain such provisos concerning native speaker adjustment and need to get clarification.

Table 6.2: Misfit with Negative Concepts

Descriptor	INFIT	Standardised (Std)	OUTFIT	Std
Needs frequently to ask for repetition, reformulation and the explanation of unfamiliar terms in order to be able to understand.	2.4	9	2.3	9
Native speakers need to make a con-	2.7	9	2.7	9

scious effort to simplify their language, keeping utterances short, using very restricted, simple vocabulary, speaking slowly and deliberately and sometimes giving exaggerated stress to key words in order for him/her to understand.

For example, rather than stating: *Needs sometimes to ask for repetition of particular words and phrases* descriptor No 215 (INFIT & OUTFIT = 1.2) was formulated: *Can follow clearly articulated speech directed at him/her in everyday conversation, though will sometimes have to ask for repetition of particular words and phrases* whilst No 138 (INFIT & OUTFIT 0.9) put the same point as follows: *Can generally understand clear, standard speech on familiar matters directed at him, provided he/she can ask for repetition or reformulation from time to time*.

In the same vein, rather than stating: *Native speakers need to articulate very slowly and carefully, with long pauses for the learner to assimilate meaning* No 24 was formulated: *Can follow speech which is very slow and carefully articulated, with long pauses for him/her to assimilate meaning*. In addition, two items which had been reworded to express Need for Support in a positive way were also successful, though the first item "overfits" considerably. Overfit means that the item is too predictable, too good to be true: the Rasch model does not quite really believe what it is seeing.

> *Can make him/herself understood and exchange ideas and information on familiar topics in predictable everyday situations, provided the other person helps if necessary.* (FIT: 0.4; Model expects 1.0, so overfitting)

> *Can interact in a simple way but communication is totally dependent on repetition at a slower rate of speech, rephrasing and repair.* (FIT: 1.4; Model expects 1.0 so rather "noisy")

This seems to reinforce the point that it is not necessarily the inclusion of negative information which is itself the problem, but rather the way in which it is used. When negative information is included in a descriptor as a qualifier in order to give information about limiting conditions or the degree of quality, this appears to function satisfactorily. In this respect it is interesting that the last descriptor, where the positive formulation is so general and vague as to be almost redundant, the technique "works" but the item is very close to the conventional criterion for misfit (1.5). The previous item is the opposite: here the descriptor could stand alone quite adequately without the proviso and the addition of the almost redundant proviso could even be what overloads the item into its "goody-goody" overfit.

The INFIT and OUTFIT statistics are usually identical. This occurred due to the corrections made to the data which were discussed in Chapter 5. For this reason only one fit statistic will be reported from now on.

Pronunciation

Pronunciation is another area which can involve an implicit negative concept: that of accent. Less accent is good, more accent is bad. Blatantly negative items showed immediately very high misfit, for example:

> *Has a strong accent which, at times, impedes understanding.* (FIT: 2.4; Standardised: 9)

However, the fact that negative wording was not the only cause of problems with descriptors for Pronunciation was indicated by the fact that

the following positively worded, functionally-oriented item misfitted equally badly.

> *Can use stress and intonation to distinguish questions from statements, and orders from requests.* (FIT: 2.4; Standardised: 9)

For this reason Pronunciation was also removed from the analysis in the early stages.

After the completion of the main analysis, supplementary analyses were then undertaken including just one of each of these excluded content strands, or more promising descriptors from them, in order to see if they would now "fit" with the construct made up by the successfully calibrated items, which were all anchored to their calibrated values.

Table 6.3: Descriptors for Pronunciation

Logit	No	Descriptor	Source	FIT	Std
3.32	274	Can vary intonation and place sentence stress correctly in order to express finer shades of meaning.	Got6 / carr7 / NewE	1.1	0

2.53	275	Has acquired a clear, natural, pronunciation and intonation.	Llb3 / natc9 / NewE	1.3	1
0.20	238	Pronunciation is clearly intelligible even if a foreign accent is sometimes evident and occasional mispronunciations occur.	CCSE3 / EC9 / RSA3 / FSI3/ sho3 / ILR3 / ESU6)	1.5	3
-2.69	160	Stress and intonation are very foreign, but can be followed okay nearly all the time.	(EC4/elviri3)	2.5	8
-3.12	108	Pronunciation is very foreign, but is clear and comprehensible within and near his/her rehearsed repertoire.	(EC3edited)	2.1	5

The analyses for each of the content strands discussed in the last section simply failed to converge as the program could not find a pattern of predictable responses which fitted with the pattern produced by the main construct and then refine it. For Pronunciation, in an analysis for the 5 more positively worded descriptors (that is still excluding the two which had misfitted dramatically early in the analysis, results shown in Table 6.3 were obtained.

The results were interesting. The 5 descriptors are calibrated in exactly the rank order expected following the order indicated by the intentions of the authors of the source scales. The bottom two descriptors are calibrated at a level later identified as *Waystage*, which also seems reasonable (though very slightly lower than the authors' intention). There is also a clear construct: (1) Before you can be followed there is not a lot to say. (2) It is easier to follow you at first when you use things you have practised—or things similar to what you have practised. (3) We can follow you but your stress and intonation are very foreign. (4) We have absolutely no problem following, you are clearly intelligible even if you have an accent and make mis-

takes. (5) You have a clear natural pronunciation. (6) You can exploit phonological features to emphasis what you want to and to communicate subtle meanings.

However, when one looks at the misfit statistics there is a clear cut-off in the scale: misfit starts with the introduction of the negative statements in No 238. This is on the conventional borderline for acceptability. Misfit of 1.5 is not excessive noisiness, and one can live with a few descriptors with a standardised residual of 3. But the bottom two items are clearly behaving in a peculiar fashion. The plausibility of No 160, with a standardised residual of 8, is very thin indeed. One could interpret this result in two ways. Firstly, one could say that in this particular case, Pronunciation, I failed to avoid the pitfall of negatively wording lower levels, so unfortunately apparent in many scales of language proficiency as pointed out by Trim (1978: 33). This may have confused the teachers, who want to say positive things about their learners and showed a distinct preference for positively worded descriptors in the pre-testing workshops. Alternatively or additionally, one could say that this has shown, once again, what a difficult thing Pronunciation is to scale since it is inevitably "normed" around the extent to which you are intelligible. At lower levels, before learners have acquired much experience of real communication in the language (and perhaps discovered the effects of being unintelligible) some people are naturally more intelligible than others,

almost irrespective of language proficiency. Pronunciation, how you sound, is an absolutely fundamental attribute of animal identity: it is how animals recognise one another (Guiora 1982: 171–6). Add to that all the personality complexes humans have about identity, convergence or divergence to group norms, accommodation theory: Giles et al 1991, Street and Giles 1982) depending on their self confidence and their attitude to the target group (socio-educational theory of SLA: Gardener 1985; Clément Gardner & Smythe 1980, Clément 1986) and it is not surprising that while some people are gifted mimics, others find it very difficult to alter pronunciation after confirming their identify at adolescence (Larsen-Freeman & Long 1991: 190). Therefore Pronunciation may be multidimensional below a point like the misfit cut-off on the scale above, and therefore unscalable there.

Suspicions about Other Content Strands

The main focus of the survey was on Spoken Interaction, which meant that two particular content strands were less central to the main construct. These two content strands were Writing (both written interaction and written production), and Spoken Production, from which Formal Presentations had already been excluded.

Writing. During the different processes in the analysis described in this chapter, the few descriptors on written production included at higher levels, plus in fact all other Writing descriptors at higher levels and all four items on Formal Correspondence were excluded as a result of misfit, instability between levels or excessive variation in the interpretation by teachers in different educational sectors or a combination between the three. The only Writing items which were not rejected during the analysis were the following lower level descriptors on clearly interactive writing (notes, forms, postcards, personal letters) which are almost spoken language written down. In other words these “written speech” descriptors fitted a construct defined by the approximately 200 items on spoken interaction.

The descriptors in the Table 6.4, in rank order by difficulty, fall into three groups. At first (bottom of table) the learner can do little more than fill in uncomplicated registration forms (the written equivalent of answering questions about personal details), and send a short simple postcard.

Then comes a band when the learner can write short simple notes, and finally, at the top, but still only at the level later identified as approximately *Threshold*, the learner can write messages or personal letters explaining and describing things in some detail. All this can be done without much knowledge of the discourse and socio-cultural conventions of the written language.

The only item with substantial misfit in No 75. Yes this seemed sensibly calibrated, and was retained since it was one of the few ALTE items left.

Table 6.4: Descriptors for Interactive Writing

Logit	No	Descriptor	Source	FIT	Std
-0.07	229	Can take messages communicating enquiries, explaining problems.	EC5	1.2	0

-0.71	239	Can write personal letters describing experiences, feelings and events in detail.	EC8	0.9	0
-1.85	75	Can write very simple personal letters expressing thanks and apology.	ALTE1 / EC2	1.9	4
-2.27	109	Can write short, simple notes and messages relating to matters in areas of immediate need.	EC3 / ASLPR1-	1.5	2
-3.28	74	Can write simple notes to friends.	EC 2–3	0.9	0
-4.04	73	Can write numbers and dates, own name, nationality, address, age, date of birth or arrival in the country etc. such as on a hotel registration form.	ILR0+ / ASLPR0+	1.5	2
-4.56	149	Can fill in uncomplicated forms with personal details name, address, nationality, marital status.	ASLPR1	1.3	1
-4.59	35	Can write a short simple postcard.	EC1	0.9	0
-4.62	34	Can fill in very simple registration forms with basic personal details.	North1 / ASLPR0+ / EC1	1.1	0

The fact that these kind of items fit so well would seem to offer some evidence for the view that this kind of low level interactive writing in the sense of "written speech" belongs under Interaction, as in the Council of Europe Common Framework, and not under Production as in the earlier version of this model (North 1992a).

Describing & Narrating (Spoken monologue). The other content strand which looked as if it could be problematic was Production. The main construct was Spoken Interaction, and interactive writing items which could be interpreted as "written speech" appeared to fit the construct, whereas items which described sustained, coherent written language *production* (for example writing reports and essays) did not. The question was whether spoken *production* would fit with Spoken Interaction. That there might be some doubt was indicated by the substantial misfit associated with Formal Presentations, one of the three categories under spoken production. The question concerned the surviving two categories Putting a Case, and Describing and Narrating.

The six descriptors for Putting a Case appeared very sensibly calibrated with exemplary fit statistics. For example No 254: *Can develop an argument giving reasons in support of or against a particular point of view* and No 213 *Can construct a chain of reasoned argument* were calibrated at a level thought to be approximately that of Cambridge First Certificate and had virtually perfect fit (0.91 and 0.95 respectively: Model expects 1.0). If there was any problem at all it was one of overfit, with a slightly easier item edited from the English National Curriculum *Can explain a viewpoint on a topical issue giving the advantages and disadvantages of various options* (FIT: 0.58).

For Describing and Narrating, although the fit appeared perfectly adequate, the issue was not so clear cut since one or two of the calibrations seemed at first sight to be rather odd. Odd calibration, as well as misfit, can be a sign that a content strand does not fit the main construct. This is why Bejar (1980) proposes that in cases of doubt, a content strand should be analysed on its own to see if there is any significant different in the difficulty estimates for the items in the content strand when they are calibrated alone or in the context of the rest of the items.

Two descriptors in particular, alternative formulations of the same idea edited from Level 2 on the ASLPR came out considerable lower than the authors had intended. The two descriptors were:

> *Can give an extended description of everyday aspects of his environment e.g. people, places, a job or study experience.*
>
> *Can describe their family, living conditions, educational background, present or most recent job.*

In Table 6.5 these two descriptors come below *plans and arrangements* and *pets and possessions*, each of which might be thought of as elementary tasks.

Table 6.5: Descriptors for Describing and Narrating

Logit	No	Descriptor	Source	FIT	Std
-1.5	66	Can give short, basic descriptions of events and activities.	North2	0.6	-2
-1.65	21	Can describe pets and possessions.	EurLonA–B / carr4edited	1.3	1
-1.93	100	Can describe plans and arrangements.	EC2 / EurLonA–B	0.7	-2
-2	40	Can describe habits and routines.	EC2 / EurLonA–B	0.7	-2
-2.01	173	Can describe past activities and personal experiences.	EC3–5 / EurLonB	0.7	-1

-2.01	209	Can give an extended description of everyday aspects of his environment e.g. people, places, a job or study experience.	ASLPR2 / EC6–7	1.1	0
-2.44	99	Can describe people, places and possessions in simple terms.	EC3 / Eur-LonA–Bedited	1.1	0
-2.51	65	Can use simple language to describe people's appear-ance.	EC2–3 / EurLonA / B / dutch1 / Lon2	1.0	0
-2.65	19	Can describe themselves and other people.	EC1	0.8	-1
-2.66	153	Can describe their family, living conditions, educational background, present or most recent job.	ASLPR2edited	0.8	-1
-3.64	20	Can describe where they live.	EC1	1.4	2

However, analysis of Describing and Narrating on its own and with the rest of production produced the same rank order of descriptors as for the main analysis. The calibrations did not change significantly, Bejar's (1980)

criterion for unidimensionality, nor did the scale lengthen significantly, Linacre et al's (1991: 3) criterion, nor was there any striking misfit. The descriptors in question share redundancy with those from other questionnaires calibrated adjacent to them. All this suggested that there was no problem of dimensionality here and that the calibrations reflected what teachers in Switzerland feel to be the case. They seem to have chosen either to ignore the word "extended" in No 209 and focus on the content, or at least to interpret it differently to the way the authors intended. The fact that *describing pets and possessions*, and *describing events and activitie*s came out higher than expected may well reflect the fact that the vocabulary needed to perform these tasks with any degree of adequacy is more extensive than is often considered to be the case. On the other hand, describing everyday aspects of one's environment like people, places, a job or study experience, one's background etc. are closer to the kind of personal details one fills out on forms and tends to comprise the content of a basic repertoire which is therefore perhaps not so difficult after all.

The result of the exclusion of problematic content strands was to clarify the construct for which descriptors could be calibrated as being predominantly informal spoken interaction, extending on the one hand to include "informal written speech" but not the written production of e.g. articles or essays, and extending on the other hand to include spoken monologue when this entailed the sustained production of a long turn embedded within an interactive context, but not to the giving of formal presentations. Within that construct, content strands involving negative concepts (e.g. need for interlocutor speech adjustment, pronunciation), and concepts which were too far removed from teacher experience (e.g. telephoning) were also excluded.

These exclusions appear intuitively sensible and can be seen as part of an overall process of refinement and exclusion which had started in the workshops with teachers. In other words participants tended to reject negative wordings and tended to be less comfortable with descriptors for sociocultural competence and work related areas. The process can be seen as continuing during the quality control on anchors and the stability of in-

dividual items described in the next two sections during which negatively worded items and the surviving items concerned with the world of work tended to draw attention to themselves, and items in content strands which could also be thought of as rather less central to the concept of proficiency in spoken interaction (e.g. strategies, listening, and above all listening strategies) were looked at with some care.

Refining the Dimension: Quality Control of Anchor Items

Having now honed down the items to what appeared to be a psychometrically adequate dimension, the next step was to refine the anchoring of the questionnaires along it.

The principle of an itembanking data collection design is that the items shared between forms (the common items, anchor items) determine the distance between the difficulty values for the different test/questionnaire forms. As mentioned in the discussion of disjunct equating versus one-step/concurrent equating Jones demonstrates that it is the anchor items common to adjacent forms which "push" the forms apart to build the dimension. One of the conditions for a successful Rasch analysis mentioned by Engelhard and Osberg (1983: 291) is that forms must be adequately and consistently linked throughout. Therefore if the anchors are badly chosen, psychometrically multidimensional, not behaving consistently but pulling in

different directions, they will distort the whole analysis. If anchor items are pulling in directions other than the one intended, the effect is to reduce the average difference in difficulty between the two forms concerned, which increases the amount of overlap between them, shortens the length of the overall scale, and thus gives lower separability and classical reliability.

Identifying Unstable Anchors

Thus it is important to establish that anchors are working as they should, and Wright and Stone (1979: 92–96) offer a quality control technique which can be used to do so. Wright and Stone suggest that the values of the anchor items on the two different forms be plotted against each other, the one on the X axis, the other on the Y axis. Then a line is drawn at the 45 degree diagonal, representing a perfect regression line. Sample anchor items are taken at different points on the dimension, the standard errors on the each of the two forms is squared, with the square root of the combination being multiplied by 1.96 to give a combined standard error expressed as a standard score. A line is then drawn at 90% from the regression line towards the anchor in question for the length of this calculated value. When several such points along the continuum are plotted, they can be linked by a line which is then a "quality control line." Any anchor occurring on the scatter plot outside that line fails a test of significance at the 95% level (or 0.05 level) and should be rejected as unstable. Such scatter plots are known as "Standard Error Plots."

In Wright and Stone's example, the 45 degree line starts at the point where the X and Y axis join as they are talking about horizontal equating: comparing two tests in the same ability range. In a vertical equating design, a more accurate picture can be obtained by taking account of the fact there is a difference of level between the two forms. This can be done by plotting the values of the higher form consistently on, say, the Y axis, and starting the 45 degree line at a point on the X axis which is the same distance away from the corner where the X and Y axis meet as the average difference in

difficulty between the set of anchors on the two forms, which is what is pushing them apart.

Each time this operation is repeated and unstable anchors are removed, the linearity of the relationship is strengthened as less "noise" is confusing the situation:

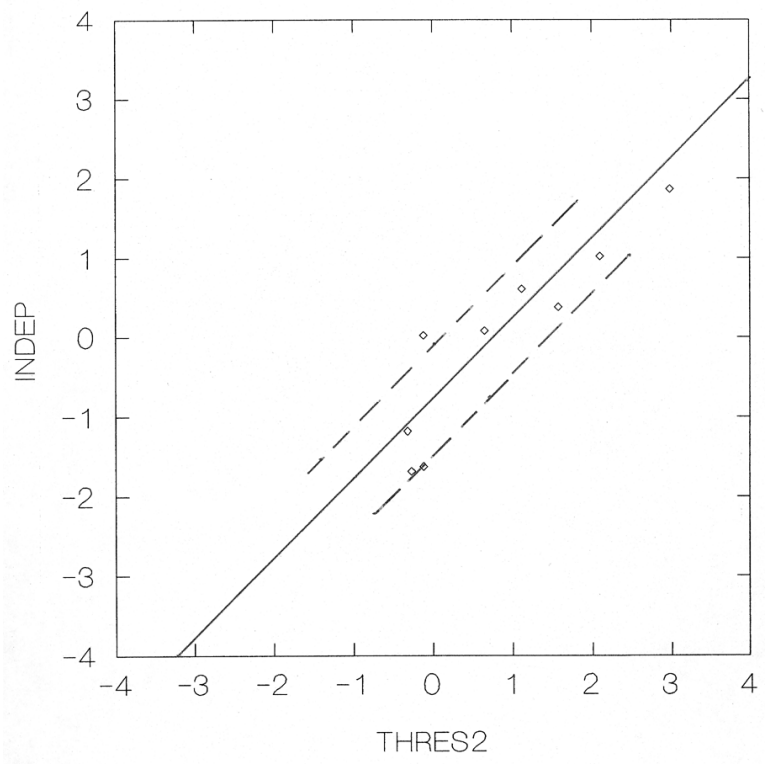
- The bulk of the remaining anchors move closer to the 45 degree line.
- One (or two) more anchors may then identify themselves as doing something different and move over the 95% criterion line, inviting exclusion from the next round.
- The average difference in difficulty between the set of anchors changes (usually lengthening), which increases the overall scale length of the full data set and improves separability.

Sometimes the operation may prove to have been unnecessary, sometimes it may only need to be repeated once, but in other cases, it may show up serious problems in the anchoring design and lead to a severe reduction in the number of anchors left. The process can be followed in the refinement of the anchoring between the Questionnaire Threshold 2 (the 5th one) and Questionnaire Independence, (the 6th one).

As can be seen in Figure 6.5, the majority of the anchor items are spread along within the criterion lines.

However, the one point outside the 95% criterion, almost exactly in the middle of the graph at approximately zero on both forms, actually represents two items placed identically. These two descriptors are not contributing to "pushing" the forms apart because they are interpreted at almost exactly the same difficulty in the environments of each of the forms, although there is now an average difference of difficulty of 0.80 logits including both these items.

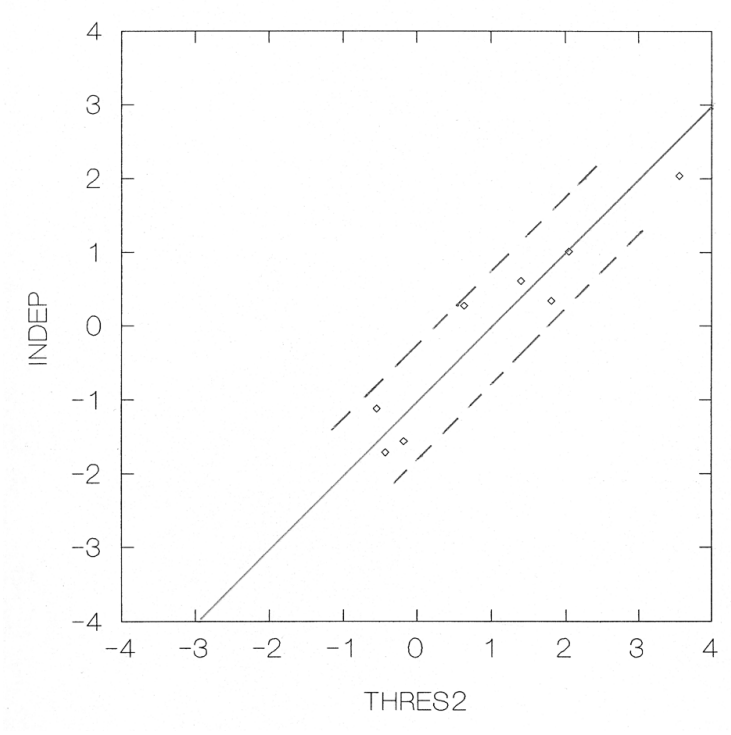
Figure 6.5: Identifying Unstable Anchor Items



The two descriptors in question are shown in Table 6.6. Notice that both are showing virtually perfect fit on both forms. Their removal gives a cleaner, less noisy vertical dimension leading to a further increase in the average difference of difficulty for the set of anchors on the two forms from 0.80 logits to 1.05. This increase produces a significant decrease in overlap between the items and persons on the two forms, more intuitively sensible calibrations and a contribution to an increase in overall scale length, separability and so reliability. As Figure 6.6 shows, after the removal of these items, the two questionnaire forms are finally linked by 8 items all of which are within the 95% criterion line. That is to say once the average difference in difficulty between the two groups of items is taken into account (by the 45 degree diagonal being moved inwards by the average difference of difficulty between the two groups of anchors) the different in values estimated on the two questionnaires for each anchor item falls with 95% certainty within the margin of standard error involved.

In other words, the difference of estimated difficulty between the two forms (adjusted for true difference of difficulty) is insignificant at the 0.05 level: the anchors can be considered stable.

Figure 6.6: Final Anchoring between two Questionnaires



Ensuring this stability had its price, however. A total of 13 anchor items failed this test of stability of interpretation at different levels. It was a shame that so many anchor items were lost, because some of them had appeared to be very nice items. Nine were the descriptors for interaction strategies, compensating strategies and listening strategies shown in Table 6.7.

Table 6.6: Unstable Anchor Items Linking Questionnaires T2 & I

		T2	I
No	Descriptor	FIT/Std	FIT/Std

194	Can maintain interaction with an interviewer, capable of responding and taking some initiatives without assistance provided there is a degree of tolerance towards lack of fluency and unconventional expression.	0.9/0	0.9/0
198	Can almost always find ways of saying what he/she wants to, with circumlocutions and some groping for words.	0.9/0	1.1/0

Table 6.7: Unstable Anchor Items Describing Strategies

No	Interaction & Compensating Strategies	Quests
41	Can ask someone to give more information.	B/W1
43	Can ask how to say an mother tongue word in the foreign language.	B/W1
44	Can identify words which sounds as if they might be "international," and try them.	B/W1
82	Can occasionally clarify what s/he means by using simple circumlocutions and extralinguistic means.	W1/W2
121	Is conscious of when he/she mixes up tenses, words and expressions, and tries to self correct.	W2/T1

198	Can almost always find ways of saying what he/she wants to, with circumlocutions and some groping for words.	T2/I
	Listening Strategies	
49	Can use the situational context to guess meaning.	B/W1
155	Can understand key words and phrases in conversations between native speakers and use them to follow the topic.	T1/T2
234	Can sustain listening and use contextual clues to identify and confirm the meaning of unfamiliar words or phrases.	I/E

It was largely because these had been considered to be interesting items that they had been included at different levels (i.e. used as anchors) in the first place. They all showed low misfit, and could have been calibrated with their ratings on just one of the two questionnaires. However, this would have been "cheating."

The fact that so many of the unstable anchors concerned strategies cannot be accidental. It in fact suggests that at least some aspects of Strategic Competence, as already suspected for Socio-cultural Competence, may form a construct which is independent of language proficiency. To calibrate items on strategies less connected to language level in this way, perhaps one should administer items to learners irrespective of language level, with learners at all levels being assessed on each form (a matrix design). However, in calibrating Strategic Competence separately, one would be left with the problem of how to equate the separate logit scales produced by the two analyses. For this reason it may be preferable to have items on Strategic Competence calibrated on the same scale as the other items, as is the case with the 29 items on Strategic Competence successfully calibrated.

Another way to look at the situation would be to say that with the exception of No 198, the Strategic Competence items in Table 6.7 all concern things that can be done to varying degrees at any level, and that the 29 items calibrated, being a little more precise, are more useful in identifying objectives for particular levels of achievement. No 198 is perhaps unstable because it is vague. It can be interpreted loosely to mean that you can *express*

the main point you want to make comprehensibly (No 182: Logit -0.98) or it can be interpreted more strictly to mean that you *can express yourself clearly and without much sigh of having to restrict what you want to say* (No 270: Logit: 2.04). No 198, being ambiguous, has been shown to be unstable.

Four other anchor items shown in Table 6.8 were also shown to be unstable. Nos 122 and 194 have probably failed for the same reasons as the strategic items. "Initiative" (194) is somewhat language-independent. It is often confused in descriptors with Turntaking. "Initiative" is in any case difficult if not impossible to display in an interview, in which one party (not you) has the right to ask questions, nominate topic and dominate the discourse (See Van Lier 1989, Silverman 1976 and a range of interaction studies at ELR Birmingham in the 1980s reported in MALS Journal).

Coping with unpredictability (No 122) is put down as Flexibility, under Pragmatic Competence, but this kind of competence could equally be seen as a Production Strategy: applying what you have to the situation in front of you. No 122 probably fails while others in this category succeed because it is so jargon-ridden and vague, a feature of many of the descriptors in the British National Language Standards from which it comes.

Table 6.8: Other Unstable Anchor Items Eliminated

No	Interaction & Compensating Strategies	Quests
194	Can maintain interaction with an interviewer, capable of responding and taking some initiatives without assistance provided there is a degree of tolerance towards lack of fluency and unconventional expression.	T2/I
122	Can cope with unpredictable elements in familiar situations.	W2/T1
45	Can manage comprehensible phrases with some effort, false starts and repetition.	B/W1
47	Shows a limited mastery of a few simple grammatical structures and sentence patterns.	B/W1

It is more difficult to see why Nos 45 (on Fluency) and No 47 (on Accuracy) should have failed to show stability of interpretation. Perhaps the fact that they are saying very little means that they fail to provide the "definiteness" of a good descriptor stated by Thorndike (1904/1912: 5 cited in Engelhard 1991a) to be essential for a valid scale. Perhaps, on the other hand, Fluency and Accuracy have a more complex relationship to each other at Beginner/Elementary level. It is also quite possible that the implicit negative concept led to inconsistent interpretation, as was the case with the misfitting items for Independence discussed in the previous section.

But it seems one can, in retrospect, see why the majority of these items are interpreted in an unstable fashion, and agree that it is correct to exclude them, although they do not show misfit on individual questionnaire forms.

Creating the Common Scale

For each final set of anchor items, two figures were calculated. The first of these was the average difference of difficulty between the set of anchors on each of the questionnaires; this was to be used to "push" the questionnaires apart (Average Pushing Factor). This gave the Average Pushing Factors shown in Table 6.9. All non-anchor items on Questionnaire I then had the value produced in their separate analysis reduced by 1.50, those on

Questionnaire T2 had theirs reduced by 2.55, those on Questionnaire T1 by 3.21 and so on.

Secondly, it seemed only sensible that the calibration of each individual anchor item should be determined by its difficulty on both questionnaires, and the extent to which this was different to the average difference.

Table 6.9: Average Pushing Factor Between Questionnaires

Questionnaires	Av. Push Factor	Cumulative
I / E	1.50	1.50
T2 / I	1.05	2.55
T1 / T2	0.66	3.21
W2 / T1	0.67	3.88
W1 / W2	0.43	4.31
B / W1	0.72	5.03

Half the difference between the average difference for the whole set, and the actual difference for this item, was therefore added to the Average Pushing Factor for the two relevant questionnaires as is illustrated for the anchors linking Questionnaires B and W1 in Table 6.10.

Table 6.10: Refining the Difficulty Values for Anchor Items

1 T2	2 I	3 Diff	4 Aver	5 Diff to Av	6 Adj1	7 Adj on Q	8 Adj2	No
-.18	-1.56	1.38	1.05	0.33	-0.17	-1.50	-1.67	191
-.43	-1.71	1.28	1.05	0.23	-0.12	-1.50	-1.62	192
3.56	2.04	1.52	1.05	0.47	-0.23	-1.50	-1.73	193
1.81	.34	1.47	1.05	0.42	-0.21	-1.50	-1.71	195
.64	.27	0.37	1.05	-0.68	0.34	-1.50	-1.16	196
1.40	.61	0.79	1.05	-0.26	0.13	-1.50	-1.37	197
2.05	1.01	1.04	1.05	-0.01	0.01	-1.50	-1.49	199
-.55	-1.12	0.57	1.05	-0.48	0.24	-1.50	-1.26	200
AV:		1.05						

The first two columns show the logit values for the items on each of the questionnaires T2 and I. The next column “Diff” gives the difference between the two values, with an average calculated at the bottom of the column. The 4th column “Aver” repeats this average and then the column “Diff to Av” gives the difference between the two figures. That figure is then halved in the 6th column “Adj1” to give the adjustment which should be made to the value of the anchor item on the higher questionnaire. This is because the value on the higher questionnaires rather than that on the lower questionnaires was used to calculate the values of the anchor items since Questionnaire E was analysed first. The 7th column “Adj on Q” gives the adjustment applied to all other items on the higher questionnaire to link its items to scale for the questionnaire above.

This is in effect the Average Pushing Factor between Questionnaires E and I as discussed above. The 8th column “Adj2” is the addition of the 6th and 7th columns. This adjustment now takes into account all the infor-

mation about the item in calculating its difficulty value, thus simulating "one-step" equating.

This put all the items onto a common scale, and calibrated anchor items taking account of all the information about them. However, the scale was now centred on Questionnaire E. That is to say zero was in the middle of Questionnaire E, not in the middle of the common scale. The last step therefore was to make a simple arithmetic adjustment to centre the scale on Zero. This could be done by taking the mean or median item difficulty and subtracting it from the value for each item. Since there were more lower level items than higher level items, using the mean would not have placed the zero point at the middle of the scale of difficulty, but rather at a weighted mid-point which would in fact have been arbitrary: purely determined by the relative number of "hard" and "easy" items. Therefore the median was used. This produced a common logit scale centred on zero.

Adequacy on Anchoring

Removal of all these anchors still left a degree of anchoring between questionnaire forms within Woods and Baker's (1985) suggestion of 3–8 items linked to the adjacent form on each side of the form in question. Taking the bank of descriptors as a whole, it also met Hambleton et al's (1991)

recommendation of 20–25% anchoring. Questionnaire I for example, the second in the Table 6.11, had 6 anchors up to Questionnaire E, and 8 questionnaires down to Questionnaire T2, and these anchors comprised 39% of the questionnaire.

Investigating Variation across Sectors and Regions

Wright and Stone's quality control technique with standard error plots which was used to identify unstable anchor items as described in the last section was also used as a way of identifying descriptors which, whilst appearing to be consistently interpreted in the global analysis, were actually displaying a statistically significant difference of judged difficulty for different educational sectors or for different language regions.

Table 6.11: Adequacy of Final Anchoring

Quest	No of Items	Anchors up	Anchors down	Proportion of Anchors
E	33		6	18%
I	36	6	8	39%
T2	34	8	7	44%
T1	37	7	11	50%
W2	37	11	7	50%
W1	37	7	7	39%
B / W1	41	7		17%
TOTAL	209	53		25%

Such variation is known technically as "Differential Item Functioning" abbreviated as DIF. This time, however, one educational sector or language region was plotted on the X axis and the other on the Y axis, always comparing the school sector appropriate for that questionnaire to the adult education sector, which used all 7 questionnaires, and the German-speaking to the French-speaking language region. The 45 degree line was drawn from

the corner this time, as in Wright and Stone's example since there was no reason to anticipate a systematic difference in the interpretation of difficulty in relation to the learners concerned.

Variation (DIF) on Individual Questionnaires

When items on a questionnaire landed outside the 95% quality control line explained in the last section, this could have been due to difference of interpretation by teachers, difference in syllabuses and/or lesser appropriacy for one of the sectors. Those items showing significant variation between sectors and regions (i.e. items which fell outside the 95% criterion line and thus show variation significant at the 0.05 level) are discussed below for each questionnaire.

On Questionnaire B (*Breakthrough*), there was no significant variation by language region, but adults were judged to find it significantly easier to *use basic greetings, to greet and introduce people* and *to use gesture to clarify what they want to say*. This does not appear surprising.

On Questionnaire W1 (*Waystage* 1), *Can use simple language to describe appearance* was judged to be significantly easier for lower secondary than for adults, and significantly easier for the French-speaking region (syllabus effect?).

The statement about Linguistic Range on both this questionnaire and Questionnaire B were interpreted as relatively more difficult for adults than for lower secondary. In other words adult sector teachers were stricter in using it, perhaps because they had more to compare to. Conversely two of the statements on strategies about *using a dictionary*, and *asking how to say something in the foreign language* were both judged to be significantly more difficult for lower secondary. As with *clarifying with gesture*, teenagers appear to use strategies less naturally than adults.

Questionnaire W2 (*Waystage* 2), covering a pre-intermediate level showed by far the most variation. 12 of the 50 descriptors showed significant variation by language region, and three by sector. Adults were judged to find the rather artificial (English National Curriculum) task of *asking written interview questions prepared beforehand* significantly more difficult; lower secondary learners were judged to find it much more difficult to *write notes and messages* and the descriptor *has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics* was considered slightly more difficult for them than for adults. The main variation, however, was by region and presents the single most significant finding of variation since clusters of descriptors are involved.

Learners in the French-speaking region were judged to find significantly more difficult all the following descriptors concerned with dealing with practical everyday situations in the foreign language: *—dealing with common aspects of everyday living such as travel, lodgings, eating and shopping; —making simple transactions in shops, post offices or banks; —using public transport: buses, trains, and taxis, asking for basic information, asking and giving directions, and buying tickets; —making a complaint; —providing concrete information required in an interview/consultation (e.g. describe symptoms to a doctor) even if with limited precision.* This result may very well reflect the more traditional pedagogic style in the Francophone region.

Learners in the German-speaking region were judged to find it more difficult to *—ask and answer questions about habits; —use written interview questions practised beforehand* or, in particular *—to ask and answer questions about pastimes.*

Ask and answer questions and respond to simple statements also belongs with this group, but just makes it inside the criterion for stability. German-speakers are also judged to find it very slightly more difficult *to produce linked sentences to convey a message.*

These results are somewhat difficult to interpret. One possibility might be that teachers in the French region interpret the phrase *ask and answer* as a unit suggesting interaction (in which there is no particular suggestion that the learner takes the initiative), whilst German-speaking teachers focused on *ask* and decided that was difficult for their learners. There is some anecdotal evidence from the workshops described in Chapter 4. to support this view. On the other hand, since none of the 6 or so other *ask and answer* descriptors were involved it could be that it is the *pastimes* and *habits and routines* that caused the problem. These themes might invite a grammatical, scholastic interpretation (can use the present and past simple) or a communicative interpretation (can engage in a meaningful exchange of information about…) and it could be that one group (Francophone) tended to the scholastic view, and that the other group (Germanophone) thought more of the communicative situation.

On Questionnaire T1 (*Threshold* 1) something interesting happened with two somewhat general, rather vague descriptors about discussion. Whereas

a quite specific descriptor on discussion *can make his/her opinions and reactions understood as regards solutions to problems or practical questions of where to go, what to do* shows excellent stability across sectors and regions and very good model fit, and straightforward ones like *can express or ask for opinions, can agree and disagree politely,* and *can express belief, opinion, agreement and disagreement* are all perfectly all right, the two more generally worded descriptors *can discuss topics of interest* and *can seek and respond to opinion on familiar subjects* showed a significantly different interpretation by both regions and by sectors. They are judged to be much more difficult for German-speakers than for French-speakers, and much more difficult for adults than for apprentices. This is somewhat counter-intuitive suggesting that these two descriptors should be treated with a little caution, even if they do end up calibrated in an apparently logical way. This example of reinforces Trim's (1978) view that vagueness in a descriptor invites a variety of interpretation.

Questionnaire T2 (*Threshold* 2) showed only one example of significant variation: the one surviving descriptor on Reception Strategies *can extrapolate the meaning of occasional unknown words from the context and deduce sentence meaning provided the topic discussed is familiar*. This sole surviving Reception Strategy was interpreted as significantly more difficult for adults than it was for apprentices and Gymnasium students. It could be that the teachers of adults were taking a more "real world" interpretation of discussion on a familiar topic, whereas the two school sectors were thinking of a tape, or classroom discussion. Alternatively it could be that the school sectors train a more analytic way of going about a comprehension difficulty, though that does seem rather unlikely.

On Questionnaire I (*Independence*), which was given to some Gymnasium and to upper intermediate learners, all 3 listening comprehension items were interpreted as significantly more difficult by the French-speaking region; this result is discussed below. For sectors the only significant difference was that the descriptor *can relate the plot of a book and give his/her reactions* was inter-

preted as significantly easier by the Gymnasium sector. This seems logical, it is rather a "schooly" task, and in fact often forms the basis of the Matura/Maturité oral interview.

On Questionnaire E (*Effectiveness*), given to some Gymnasium and to advanced learners, two descriptors both dealing with sustained, coherent, spoken production *can write reliable reports on extended spoken or written information* and *can give clear detailed descriptions of complex subjects* were considered far easier for Gymnasium students than they are for adults, which again has a certain logic. There was no significant different by language region.

Conclusions of DIF

Many items which showed variability across regions or across sectors appeared nonetheless to be good items, well calibrated, well fitting, sensible, saying something. For example, on Questionnaire *Independence* (for the sake of argument, roughly First Certificate Level; Interagency Language Roundtable 2; Eurocentres 6/7), does the fact that the 3 listening comprehension items failed 95% confidence on regions—the French-speaking region considering them much more difficult than the German-speaking region—make them bad items? A recent analysis of the main course used in the French-speaking cantons concluded that the classroom practice of listening comprehension was minimal. This could be a problem of an inadequate syllabus. One of these items: *can sustain listening and use contextual clues to identify and confirm the meaning of unfamiliar words or phrases* was an anchor which failed 95% confidence once worse anchors had been removed, and also failed 95% confidence across sectors on one questionnaire form. Like most of the other items on listening strategies, it was therefore dropped. A second item: *can get the gist of most of what is said in conversation and discussion around her on topics which require no specialised knowledge*, which had been problematic in the initial analyses with a very counter-intuitive calibration, was revealed to also be noisy (1.6 or more) both globally and on sectors, and still to have a calibration which seemed lower than sensible. Since this was the opposite of what the Francophones were here saying, something odd was clearly happening so the item was dropped.

The third item, however: *can follow clearly articulated speech directed at him/her in everyday conversation, though will sometimes have to ask for repetition of particular words and phrases*, appeared on all accounts to be a perfectly adequate item. Two points present themselves:

Firstly, Listening in Interaction is arguably not so central to the construct; it can be separately identified and taught/not taught. Yet the 12 listening items which survived seem well-fitting and well-calibrated where their authors had intended. Should they therefore be excluded from a common scale because there is possibly a significant difference between regions

regarding comprehension ability? It would seem not. Different achievement by learners from the different regions can be reported on the same scale; that is the point of a common scale.

Secondly, the first two items discussed have something odd about them. The first appears to be interpreted at different levels of achievement in a way not consistent with the way the majority of items are interpreted. The difference between levels is not great enough; the anchor does not "push." This is relatively common with items on aspects of strategic competence, as we saw in the last section. Such items may not be bad items, but they may be better calibrated in a fully separate analysis, which would entail anchoring this strand securely across all questionnaire forms. However, this particular item is significantly unstable across regions and sectors, so maybe it is just too vague. With the second item, one is tempted to speculate that the problem is *gist*. Gist is an odd concept. How can you get the gist of something if you do not follow it all? If you could follow it all, then you got more than the gist, even if you can only remember the gist.

The example of variation in listening comprehension across regions was taken because listening is a fairly obvious content strand. Similar arguments can be made about variation between sectors, where for example the variation at advanced level (Questionnaire E) on sustained, coherent production tasks, (easier for academic Gymnasium students) is, as stated, only logical.

The large number of significant differences of interpretation of the difficulty of "real life tasks" on Questionnaire W2 is certainly interesting, but again it may well say more about the pedagogic culture in French-speaking Switzerland than it does about the descriptors.

Failing 95% stability across sectors /regions draws attention to an item, but is *not* in itself necessarily an argument for dropping the descriptor concerned. Twenty-five items which showed significant variation across sectors or regions, but no other signs indicating that they were problematic were retained in the descriptor bank. However, such items which show significant variation across sectors or regions should only be included in profiling grids which are meant to be used to plot such variation, and should *not* be subsumed into a holistic descriptor on a global scale purporting to report achievement independent of context. Putting this result more positively, 183 of the 209 descriptors finally calibrated (87.5%) show no significant (0.05) variation across sectors or regions.

Refining the Bank: Quality Control on Individual Descriptors

Now that the factors causing distortion in the data had been dealt with, the use of the rating scale had been checked, the construct had been honed down to what seemed likely to be sufficiently psychometrically unidimensional, the stability of the anchors had been confirmed and the degree of differential item functioning (DIF) had been investigated, the outline of the descriptor bank had been established. It remained to conduct quality control on each of the descriptors to decide which to include or exclude in order to rerun the analysis to obtain final difficulty estimates. In fact by this stage 5 individual items had already been removed due to high misfit when the questionnaires were first analysed separately with the conference data removed. The 5 items were those shown in Table 6.12.

In retrospect, it is not difficult to see why they were poor items. The first two are Strategies (Turntaking/Repairing), which have already tended to be problematic unless well worded. Both descriptors invite an interpretation in relation to non-verbal strategies, i.e. abilities other than

language proficiency. The former is also an example of a "double-barrelled" descriptor linked by a "but" which had proved unpopular in the workshops. The latter also involves a certain contradiction: will you be allowed to intervene regularly in other people's conversations if you have a habit of doing so in an inappropriate fashion? The next three are Writing items, and each in their own way on the fringe of a Writing construct.

Table 6.12: Items with High Misfit

No	Descriptor	Quest	FIT
156	Can regularly join in a conversation, but may often do so inappropriately.	T1/T2	2.3
180	Can usually clarify meanings by using circum-locutions and other repair strategies, but relies more on extralinguistic strategies than on verbal ones.	T2	2.0
76	Can add an address, date, title and pre-arranged opening and closing formulae to formal letters.	W1	2.6
189	Can take down a simple message in note form.	T2	1.9

280	Can edit and redraft, possibly using reference books, checking for accuracy and appropriacy or expression.	E	2.1

After the investigation of variation by sector, items which had shown gross misfit in one sector or in one region were also excluded. This affected a further 4 items which are given in Table 6.13.

Of these descriptors 3 are also on Writing, whilst the last combines (a) a suggestion of non-linguistic aspects (*Is at ease with*) and (b) Socio-cultural Competence, and misfits badly both on one sector and on one language region.Item performance histories were recorded for all items for which the interpretation of difficulty had shown a significant difference between sectors or regions, or which gave other grounds for suspicion like noticeable misfit (mean square 1.5) in the main analysis or a calibration which looked odd.

The reasons for logging items in the item performance histories led to the emergence of four criteria of negative quality:

- "noisiness" in global (mean-square) fit: i.e. 1.5 or above
- "extreme noisiness" in one sector or region (1.7)
- failing 95% confidence for stability across sectors or regions
- suspicious looking calibration
- anchor refusing to "push"

Table 6.13: Items with High Misfit on One Sector / Region

No	Descriptor	Quest	FIT
110	Can write simple, short formal letters by adapting part of a given model.	W2	Sector misfit: 2.16
125	Can write personal letters to a friend, host etc. giving and asking for news.	W2/T1	Sector misfit: 2.07

278	Can write clear, well-structured formal letters in an appropriate style.	E	Sector misfit: 2.60
233	Is at ease with "small talk" in most kinds of social situations, familiar with conventions of polite conversation.	I/E	Sector misfit: 2.44 Region misfit: 2.16
280	Can edit and redraft, possibly using reference books, checking for accuracy and appropriacy or expression.	E	Sector misfit: 2.1

Failing one of these criteria was not considered sufficient for removal. Some "noisy" items could be tolerated in the bank, provided the calibrations seem sensible. Misfit is a complex phenomenon, and is rarely if ever excluded completely since when misfitting items are removed, other items start to misfit. Analysis requires judgement, not just eliminating things from a computer print out on the basis of numbers alone: "fit statistics are indicative not absolute" (Linacre 1990: 7).

A total of 71 items drew attention to themselves by failing one of the four criteria listed above. Many items which in the initial analyses had appeared a little odd could now, thanks to these criteria, have that oddity ex-

plained. In some cases, it became clear why an anchor had failed to "push" (contribute to the difference of difficulty between adjacent questionnaires). If, for example, an anchor was interpreted radically differently by two sectors or regions, or misfitted badly in one sector or region, the two interpretations could in effect cancel each other out.

None of the criteria were considered sufficient grounds in themselves to reject an item since:

- In relation to "noisiness," the Rasch model is pretty robust; as was seen in the discussion of Pronunciation even very noisy and misfitting items are usually calibrated sensibly.
- In relation to "noisiness" in one sector, the fact that one sector were a little inconsistent in their use of an item which was functioning well in the other sector did not by itself seem a sufficient reason for removal.
- In relation to stability of interpretation of difficulty across sectors and regions, as discussed earlier there could be perfectly valid grounds for such differences.
- In relation to a suspicious-looking calibration. This could be a second reason for being cautious about an item which had failed one of the other criteria but was hardly a reason in itself. As discussed in relation to Describing and Narrating calibrations which at first appear to be surprising because they contradicted preconceptions could well be correct.
- Finally, in relation to anchors refusing to "push," automatic exclusion of items which whilst within the margin of error defined by the 95% criterion show little difference of difficulty on two different forms would over-weight the contribution of those items for which this difference is *over*estimated and would therefore exaggerate the difference between forms and lead to spurious scale length. Removing them when it becomes clear that the lack of "push" is not due to chance, but can be explained by one of the other criteria is another matter.

The stipulation that failure on two criteria removed an item may have been a little strict, but it seemed preferable to err on the side of caution. Detailed investigation of item histories in this way led to the exclusion of the 9 items shown in Table 6.14. Of these 9 items, 3 again concern more formal, coherent written production. As such they are very different from spoken language. Formal letters are very different to personal letters in this respect.

Table 6.14: Items Failing Two Quality Criteria

No	Descriptor	Quest	Reason 1	Reason 2
70	Can consult a dictionary to find phrases which, even if not lexi-cally appropriate, have a good chance of being comprehensible.	W1	Failed 95% significance on Sectors	Calibration okay, but different from all other items. Danger "could be any level."

78	Can ask written interview questions he/she has prepared and practised beforehand e.g. about leisure activities, food preferences.	W1 / W2	Failed 95% significance on Sectors & Regions	Artificial "schooly" task
216	Can get the gist of most of what is said in conversation and discussion around … him/her on topics that require no specialised knowledge.	I	Failed 95% significance on Regions very badly	Very strange calibration

Table 6.14 (cont.) : Items Failing Two Quality Criteria

No	Descriptor	Quest	Reason 1	Reason 2
230	Can write formal letters to order goods, book a room etc.	I	Failed 95% significance on Regions	Calibration looks questionable
240	Can write reliable reports on extended spoken or written information.	I / E	Failed 95% significance on Sectors	Written Production: not part of main construct
126	Can negotiate a price e.g. for a second hand car, bike.	T1	Failed 95% significance on Regions	Calibration looks questionable

161	Can use persuasive language to try and get a lower price or an added extra, e.g. when hiring a car, renting a room.	T2	Failed 95% significance on Regions	Calibration looks questionable. Suspect a "block effect"—loss of local independence
246	Can hold his/her own effectively in negotiations, explaining clearly the reasons for the position he/she is taking, and the limits or qualifications to possible concessions.	E	Failed 95% significance on Regions	Only "Business" item left: not central to construct

As mentioned before in this chapter, two overlapping constructs are determining the bank. These two overlapping constructs were *Interaction* including interactive writing, and *Speaking* including spoken production, which is sustained long turns in normal conversation and discussion, but

not formal presentations, which share many of the characteristics of written language. The construct in the bank has also excluded areas which could not be observed and/or were work-related (Telephoning, Meetings). Three more of the items above concern Negotiating, which seems more and more to be on the fringe of the construct. There would be an argument for excluding Negotiating altogether, but since in practice the surviving 4 items on Negotiating come out sitting on top of the scale on Service Encounters, with the top two items on Service Encounters beginning to suggest Negotiation, it seems it can best be regarded as a continuation of the Service Encounters sub-scale (See Service Encounters sub-scale in Appendix 3).

Of the other three items, Nos 70, 78 and 216, the former two could actually have been left in the bank, but on reflection they describe rather artificial school tasks of a kind not found elsewhere in the bank, which is presumably causing the variation by sector, and which is an argument for exclusion. No 216 is interesting in that it calibrated extremely oddly through all phases of the analysis and frankly, one was glad to have an excuse to get rid of it. The problem concerns the concept of "gist" which seems to suggest that somehow in listening, top-down processing is independent from bottom-up processing, rather than the two being wedded in an interactive process as for reading (Rumelhart 1977; Eskey 1988; Moran and Williams 1992). "Gist" seems a relic of the Goodman's top-down model of a "psycholinguistic guessing game" (Goodman 1967 cited in e.g. Carrell and Eisterhold 1983; Eskey 1988; Moran and Williams 1992: 65). As such, one could argue that the teachers in the German-speaking region, who interpreted this as a *very easy* item on Questionnaire I (over -4.0 logits: the easiest of the 50 items), are taking a rather optimistic Goodman-based view, whilst their French colleagues, faced with the listening comprehension difficulties that their learners at this level continue to have, are considerably more restrained (circa -1.0 logits: the 10th easiest item). An alternative explanation might be that this item falls into the same group as Telephoning, Meetings etc: these teachers do not really know about "gist," because they never ob-

serve this behaviour, and in guessing, they are over-influenced by the sway of a simplistic interpretation of receptive processing as a guessing game.

A third possibility related to the two mentioned above is that the lack of definition of "gist," the failure to identify the processes involved in it, leads to an underestimation of how difficult it is. The item on this kind of comprehension shown in Table 6.15 displayed no variation between sectors or regions, considerable overfit (i.e. very high consensus which the model is rather suspicious of) and sensible calibration at what was later associated with *Threshold Level*. This is about gist, but it explains more precisely what is meant and adds a proviso about when the learner can do it.

Table 6.15: A Good Gist Listening Item

Logit	No	Descriptor	Fit / Std	Source
-1.04	176	Can generally follow the main points of extended discussion around him/her, provided speech is clearly articulated in standard dialect.	0.25 /0	North4 / AMES3 / elviri3

Finally, 4 other items were excluded from the original 280. Firstly a duplicate which had slipped in, and secondly the last remaining item on formal writing which was somewhat "noisy" (Fit: 1.44/2) and which, when one looked at the descriptors adjacent in the calibrated rank order seemed to be being calibrated on the basis of the theme rather than the task, on the basis of "routine (…) standard phrases" rather than "write simple formal letters." The last exclusion was also perhaps overcautious and concerns an item on Independence (Need for Help) with a positive wording which seemed to have been calibrated a bit high.

Table 6.16: Other Questionable Items

Logit	No	Descriptor
-3.70	94	Can ask and answer questions and participate in short conversations in *routine contexts* on topics of interest.
-3.71	190	Can write short, simple, routine formal letters consisting mainly of standard phrases.
-3.73	83	Can adapt well rehearsed *memorised simple phrases* to particular circumstances through limited lexical substitution.
-3.86	141	Can ask for clarification about key words not understood using *stock phrases*.
	188	Can make him/herself understood, but needs an interlocutor whom he/she can ask for help if he/she tries to say exactly what she wants to.

Item 188 was temporarily excluded, and then, after the other items had been calibrated, an attempt was made to reinstate it. The analysis failed to converge, suggesting that it had been correct to regard it with suspicion.

Establishing an Item Quality Hierarchy

After all the adjustments described in Chapter 5, and the exclusions discussed earlier in the chapter the scale produced contains 212 descriptors (including the 3 reinstated Pronunciation items).

Of these calibrated descriptors, 15 show misfit (outfit mean square) of 1.50 or more and 23 show a standardised residual of 2.0 or more. Averaging between these two conventional statistical cut-off criteria, just under 10% of the descriptors show significant misfit. According to Stansfield and Kenyon (1992: 10) less than 10% of the items should be misfitting before adequate fit to the Rasch model and necessary psychometric unidimensionality can be claimed. By that criterion the full scale of 212 descriptors produced in the 1994 study is on the borderline of psychometric acceptability. Considering the real multidimensionality implied by the range of content strands, educational sectors, language regions, local education systems, and teacher experience (from 3 weeks to over 20 years), that is quite a satisfactory result.

But different purposes require different degrees of rigour. For use in teacher continuous assessment to profile achievement, looser criteria can be applied than for the production of a holistic summary scale against which to report relative achievement in different sectors and regions. Accordingly, three criteria were used to create a quality hierarchy. These three criteria were the following:

- *Mean Square Fit*: (Amount of misfit). Good fit: 0.5 to 1.5.
- *Standardised Fit*: (Plausibility). Good fit: -3 to +2. Overfit less important. Very good model fit: -1.5 to +1.5
- *Stability*: Good Stability: 95% Standard error plot criterion (Sectors / Regions), Excellent stability: Twice this standard criterion

The resulting quality hierarchy of excellent, very good and adequate items is shown in Table 6.17.

"Excellent items" display a really surprisingly high degree of consistency and stability of interpretation in different settings. Four of the five descriptors for *Fluency* came out in this category, making Fluency the most consistently interpreted and therefore most generalisable descriptor category in the survey, a finding repeated in the second year. Other categories with a concentration of "excellent items" were Cooperating Strategies and Grammatical Accuracy (each with 3 out of 6 descriptors). Cooperating Strategies was included by a number of more communicatively inclined teachers to be part of a broader concept of Fluency. The categories with the greatest concentration of highly stable, highly consistent items could thus be said to reflect the fundamental and very popular Fluency / Accuracy distinction established by Brumfit (1984). Because of their excellent fit and stability, such items are eminently suited to being used as anchor items to expand the numbers of descriptors in the bank. This is demonstrated in a study equating a scale of "Can do statements" produced by the Association of Language Testers in Europe (ALTE) to the scale produced in this study. In the equating study, carried out in late 1999, the ALTE scale values of a set of 16 such "excellent items" used as anchors correlated 0.97 to their values in this study (n = circa 1,500).

Table 6.17: An Item Quality Hierarchy

	Characteristics	Identified in Appendix 3
Excellent Items	High stability and consistency of interpretation and extremely high probability with very good model fit: 1.5 or less mean square fit; -1.5 to + 1.5 Std); tighter criterion on variation	Bold
Very Good Items	95% Confidence criterion for sectors & regions; normal model fit	Normal
Adequate Items	Good calibration; but with some variation across sectors/regions and/or "noisiness" in model fit. Some multidimensionality here.	*Italics*

"Very Good Items" are the main body of items in the descriptor bank. Because they display stability across contexts as well as good model fit, they would be suitable for inclusion in the drafting in a holistic summary scale to explain levels. They could also form the basis for the development of transparent assessment criteria for more formal assessment situations.

"Adequate Items" are suitable for inclusion in a descriptor bank from which checklists will be developed for continuous assessment and from which grids will be developed to profile performance across categories. Because of their degree of "noisiness" or instability they are inappropriate for a summary scale and should if possible not be used as anchor items since their statistical imperfection could distort estimates in a future analysis.

7 Interpreting the Scale

212 descriptors had now been calibrated to estimated difficulties on a common logit scale running from -5.68 to 4.68. Having successfully constructed a scale of items, the next step was to investigate it firstly in order to check that it indeed made sense, that similar content was calibrated in a coherent fashion, and secondly in order to present it in a form in which it could be meaningful to other users. In effect this meant dividing the scale up into a number of bands or levels, which in turn involved setting cut-off points, and then seeing (a) whether those levels had coherent content (b) whether progress up the scale in each category was logical.

Setting Cut-offs between Levels

The number of levels or strata which can be identified in a set of data is connected to the question of reliability. Pollitt explains a calculation with

Table 7.1: Reliability and the Number of Strata in Data

r	Pollitt 1991 Bands	Fisher 1992 Distinct Strata
0.98		9
0.97		8
0.96	10.1	7
0.94		5
0.90	6.3	4
0.80	4.3	3
0.70		2
0.50		1

which one can derive the decision capability from any test (or rating) from its reliability coefficient (Pollitt 1991: 90). Fisher (1992) offers similar information and his table and Pollitt's compare as in Table 7.1.

The reliability statistic for the full integrated analysis (simulated Cronbach Alpha) was 0.97, which, according to Pollitt would justify 10 bands/levels, or according to Fisher: 8 bands.

Ways of Setting Cut-offs

Setting cut-offs is notoriously subjective, as mentioned in discussing criterion-referenced assessment. In the current context, there seemed to be three possible bases for doing it:

1. By referring to logit values in an attempt to create a scale of more or less equal intervals.
2. By looking for patterns and clusters, and apparently natural gaps on the vertical scale of descriptors which might indicate "thresholds" between levels.
3. By comparing such patterns to the intentions of the authors of the source scales from which descriptors had been taken or edited, and to the posited conventional or "natural levels" (Hargreaves 1992; North 1992a).

Logit Range would appear to be the obvious place to start. Even if one has reservations about the ability of the Rasch model to produce a linear scale after all the problems encountered, it would be absurd to ignore the logit range and make no attempt to create levels or bands at equal intervals when one knows that non-specialists tend to interpret scales of proficiency as if they were in fact linear. On the other hand, precisely because there do appear to be genuine reservations about the true linearity of a logit scale, even when steps had been taken to avoid error as in this case, there seemed little point being fanatical about it if looking at the content appeared to suggest continuing scale distortion.

Clusters and Gaps on the Scale might well be an artefact of the analysis, an accident caused by what was and was not included, but on the other hand, since nearly all of the descriptors came from scales which aimed to describe distinct levels, gaps might actually indicate gaps between those source levels.

Author Intentions are a way of checking the plausibility of difficulty estimates. Wright and Masters (1982: 90) go as far as to suggest consider that calibrations must conform to author intentions before the calibrations can

be accepted as reasonable. This seems somewhat overstated, since it contradicts Thurstone's (1928b) stipulation that scale values should be independent of the opinions of the authors, or particular groups. After all, the whole point of this study was to see if the opinions of the original authors were empirically supported. One might reformulate the point to say that values should be independent of the views of the authors, but must make sense. If there is a contradiction to the views of the authors, it should be possible to explain why the analysis is right and the authors were wrong.

Approach Adopted to Setting Cut-offs

Griffin discusses looking for patterns in the primary reading scale (1989; 1990a: 297) and marking out equal intervals on the logit scale for adult literacy (1990b: 11). In this study all three methods listed above were used in order of the priority of the numbering.

Firstly equal distances on the logit scale were marked out and the boundaries were then shifted occasionally to coincide better with where there were natural gaps. Then the coherence of the content of those provisional "levels" was investigated and compared to (a) the intentions of the source scale authors and (b) conventional levels like *Threshold Level* and *Waystage*. In other words, the question was asked: does this content seem to fit together and can it be labelled in its own right and/or in relation to con-

ventional levels? Finally the equality of the intervals on the scale was re-examined.

Equal Interval Levels and Common Reference Levels

In this way a set of levels of approximately equal intervals was established. Each level covers approximately 1 logit, and the fact that attention to content coherence leads the range to be slightly narrower in the middle of the scale (0.97/0.98 logits) and wider at the ends (1.10 logits) could be regarded as reflecting the tendency for a Rasch logit scale to distort towards the ends, despite all the corrective measures taken as described in Chapter 5. The levels have been given names related to (a) the conventional levels (*Breakthrough, Waystage, Threshold, Independence, Effectiveness, Mastery*) which were used to organise the descriptor pool and then to construct questionnaires and (b) the name of a new level specification (*Vantage*) since written for the Council of Europe (Van Ek and Trim 1995). The third column in Table 7.2 gives the original names used in this study, the columns to the left, the labels used in Council of Europe Common European Framework (Council of Europe 1996).

Table 7.2: Equal Interval Levels and Common Reference Levels

Common Reference Levels		Finer Level (Swiss)	Abbrev	Cut-off	Range on Scale
C2	Mastery	Mastery	M	3.90	
C1	Effective Operational Proficiency	Full Effectiveness	EOP	2.80	1.10
	(*Vantage Plus*)	Effectiveness	V+	1.74	1.06
B2	Vantage	Full Independence	V	0.72	1.02
	(*Threshold Plus*)	Independence	T+	-0.26	0.98
B1	Threshold	Threshold	T	-1.23	0.97
	(*Waystage Plus*)	Waystage Plus	W+	-2.21	0.98
A2	Waystage	Waystage	W	-3.23	1.02
A1	Breakthrough	Breakthrough	B	-4.29	1.06
	---------	Tourist	Tour	-5.39	1.10
	--------	Smattering			

Smattering is not a level, but appears rather to be characterised by the ability to perform certain very simple, isolated tasks: e.g. *use some basic greetings* and *say yes, no, excuse me, please, thank you, sorry.*

Tourist is so labelled because the 5 descriptors calibrated there —*fill out registration forms; write a postcard; ask and tell time; make simple purchases* are reminiscent of the things one does as a tourist. It represents the lowest level that is likely to be useful as an objective but seems also to characterise the ability to perform particular isolated tasks rather than a level of generative language use. Subsuming "*Tourist*" into *Breakthrough* (a level used in LANGCRED), gives 9 levels, mid way between Pollitt's and Fisher's interpretations.

Breakthrough was so named by LANGCRED to give a name to what they considered the lowest level of generative language use, the point at which the learner can: *interact in a simple way; ask and answer simple questions about themselves, where they live, people they know, and things they have; initiate and respond to simple statements in areas of immediate need or on very familiar topics*, rather than relying purely on a very finite rehearsed, lexically organised repertoire of situation-specific phrases, as is probably the case at *Tourist.* It is Level A1.

Waystage appears to reflect the elementary level referred to by the Council of Europe specification. It is at this level that the majority of descriptors stating social functions are to be found, like: *use simple everyday polite forms of greeting and address; greet people, ask how they are and react to news; handle very short social exchanges*; *ask and answer questions about what they do at work and in free time; make and respond to invitations; discuss what to do, where to go and make arrangements to meet*; *make and accept offers.* Here too are to be found descriptors on getting out and about reflecting the simplified cut-down version of the full set of transactional specifications in *Threshold* for adults living abroad, like: *make simple transactions in shops, post offices or banks; get simple information about travel; use public transport: buses, trains, and taxis; ask for basic information; ask and give directions, and buy tickets; ask for and provide everyday goods and services.* It is Level A2.

Waystage Plus is the name given to a band representing a weaker realisation of Threshold content, or strong Waystage performance. It is identified as Level A2+ or A2.2 in the Common Framework. What is noticeable here is more active participation in conversation given some assistance and certain limitations, for example: *initiate, maintain and close simple, restricted face-to-face conversation; understand enough to manage simple, routine exchanges without undue effort; make him/herself understood and exchange ideas and information on familiar topics in predictable everyday situations, provided the other person helps if necessary; communicate successfully on basic themes if he/she can ask for help to express what he wants to; deal with everyday situations with predictable content, though he/she will generally have to compromise the message and search for words; interact with reasonable ease in structured situations, given some help, but participation in open discussion is fairly restricted* and more ability to sustain monologues, for example: *express how he feels in simple terms; give an extended description of everyday aspects of his environment e.g. people, places, a job or study experience; describe past activities and personal experiences; describe habits and routines; describe plans and arrangements; explain what he/she likes or dislikes about something; give short, basic descriptions of events and activities; describe pets and possessions; use simple descriptive language to make brief statements about and compare objects and possessions*.

Threshold is intended to represent the Council of Europe specification for a visitor to a foreign country and is Level B1. It is perhaps most categorised by two features. Firstly there is an ability to maintain interaction and get across what you want to in a range of contexts, for example: *generally follow the main points of extended discussion around him/her, provided speech is clearly articu-*

lated in standard dialect; give or seek personal views and opinions in an informal discussion with friends; express the main point he/she wants to make comprehensibly; exploit a wide range of simple language flexibly to express much of what he or she wants to; maintain a conversation or discussion but may sometimes be difficult to follow when trying to say exactly what he/she would like to; keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production. Secondly there is an ability to cope flexibly with problems in everyday life, for example *cope with less routine situations on public transport; deal with most situations likely to arise when making travel arrangements through an agent or when actually travelling; enter unprepared into conversations on familiar topics; make a complaint; take some initiatives in an interview/consultation (e.g. to bring up a new subject) but is very dependent on interviewer in the interaction; ask someone to clarify or elaborate what they have just said.*

Independence was the name used in the study to describe the questionnaire above Threshold. The term Independent User was employed for a time by UCLES to describe the level of First Certificate and was adopted by LANGCRED to describe *Threshold*. It was found useful during questionnaire preparation but is replaced here partly because it seems to be applicable to a broad range rather than a narrow range of level (c.f. different uses by UCLES and LANGCRED) and partly because of the emergence of the

name *Vantage* (Van Ek and Trim 1996) to describe a level which seems to represent a new "threshold." The content of this level previously referred to as *Independence* seems actually to be a strong version of *Threshold*, hence the adoption of *Threshold Plus*. In the Common Framework it is referred to as Level B1+ or B1.2. The same two concepts as at *Threshold* continue to be present, with the addition of a number of descriptors which focus on the exchange of quantities of detailed information, for example: *take messages communicating enquiries, explaining problems; provide concrete information required in an interview/consultation (e.g. describe symptoms to a doctor) but does so with limited precision; explain why something is a problem; summarise and give his or her opinion about a short story, article, talk, discussion interview, or documentary and answer further questions of detail; carry out a prepared interview, checking and confirming information, though he/she may occasionally has to ask for repetition if the other person's response is rapid or extended; describe how to do something, giving detailed instructions; exchange accumulated factual information on familiar routine and non-routine matters within his field with some confidence.*

Vantage, originally labelled *Full Independence*, does appear to represent a significant shift. This offers some justification for the new name *Vantage*, (Level B2) intended for a new level as far above *Threshold* as *Waystage* is below it. According to Trim (personal communication) the intention is, as with *Threshold* and *Waystage*, to find a name which has not been used before, and which symbolises something central to the level concerned. In this case, the metaphor is that having been progressing slowly but steadily across the intermediate plateau, the learner finds he has arrived somewhere, things look different. He/she acquires a new perspective, can look around him/her in a new way. This concept does seem to be borne out to a considerable extent by the descriptors calibrated here, which represent quite a break with the content so far. For example at the lower end of the band there is a focus on effective argument: *account for and sustain his opinions in discussion by providing relevant explanations, arguments and comments; explain a viewpoint on a topical issue giving the advantages and disadvantages of various options; construct a chain of reasoned argument; develop an argument giving reasons in support of or against a particular point of view; explain a problem and make it clear that his counterpart in a negotiation must make a concession; speculate about causes, consequences, hypothetical situations; take an active part in informal discussion in familiar contexts, commenting, putting point of view clearly, evaluating alternative proposals and making and responding to hypotheses*. Running right through the level are also two new

themes: firstly being able to more than hold your own in social discourse, for example: *converse naturally, fluently and effectively; understand in detail what is said to him/her in the standard spoken language even in a noisy environment; initiate discourse, take his turn when appropriate and end conversation when he/she needs to, though he/she may not always do this elegantly; use stock phrases (e.g. "That's a difficult question to answer") to gain time and keep the turn whilst formulating what to say; interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without imposing strain on either party; adjust to the changes of direction, style and emphasis normally found in conversation; sustain relationships with native speakers without unintentionally amusing or irritating them or requiring them to behave other than they would with a native speaker.* Secondly there is a new degree of language awareness: *correct mistakes if they have led to misunderstandings; make a note of "favourite mistakes" and consciously monitor speech for it/them; generally correct slips and errors if he becomes conscious of them; plan what is to be said and the means to say it, considering the effect on the recipient/s.* Taken together, this does seem to be a new ball game, a new threshold.

Vantage Plus was originally labelled *Effectiveness.* The descriptors with the word "effective" in them, which had been on Questionnaire E (Effectiveness) did actually land at this level: *use the language fluently, accurately and effectively on a wide range of general, academic, vocational or leisure topics; carry out an*

effective, fluent interview, departing spontaneously from prepared questions, following up and probing interesting replies. The focus on argument, effective social discourse and on language awareness which appears at *Vantage* continues, and the level is identified as B2+ or B2.2 in the Common Framework. However, the former two can also and perhaps more usefully be interpreted as a new focus on discourse skills, both in terms of conversational management (cooperating strategies): *give feedback on and follow up statements and inferences by other speakers and so help the development of the discussion; relate own contribution skilfully to those of other speakers*, and coherence/cohesion: *use a limited number of cohesive devices to link sentences together smoothly into clear, connected discourse; use a variety of linking words efficiently to mark clearly the relationships between ideas; develop an argument systematically with appropriate highlighting of significant points, and relevant supporting detail*, and a concentration of items on Negotiating: *outline a case for compensation, using persuasive language and simple arguments to demand satisfaction; state clearly the limits to a concession.*

Effective Operational Proficiency was originally labelled *Full Effectiveness*. What actually seems to characterise this level is good access to a broad range of language, which allows fluent, spontaneous communication. On the Eurocentres scale Level 9, the level below *Mastery*, is given the label "Fluent User" and descriptors focus on the appropriate use of a wide range of language. This is quite similar to the concept which is described by the 6 descriptors calibrated at this level: *can express him/herself fluently and spontaneously, almost effortlessly; has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions. There is little obvious searching for expressions or avoidance strategies; only a conceptually difficult subject can hinder a natural, smooth flow of language.* However, colleagues in both the Council of Europe and Swiss projects objected to that label "Fluent User" on the grounds that one can be fluent at any level. Therefore the name Effective Operational Proficiency was adopted. The level is referred to as C1 in the Common Framework. The aspects focused on at the level below are also in evident at this level: discourse skills, both conversational management (though again with an emphasis on fluency): *select a suitable phrase from a fluent repertoire of discourse functions to preface his remarks in order to get the floor, or to gain time and keep it whilst thinking.* and coherence/cohesion: (though again with an emphasis on smoothness, flow): *produce clear, smoothly-flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices.*

Mastery is not intended to imply native-speaker or near native-speaker competence. As argued in Chapter 2 there are grounds for considering the "educated native speaker" an inappropriate benchmark. Rather what was intended by the expression is the degree of precision and appropriateness and the ease with the language which characterises the speech of those who have been highly successful at learning the language. Only three descriptors were calibrated here in Year 1, but between them they paint a picture of precision, colloquial appropriateness and ease of expression: *convey finer shades of meaning precisely by using, with reasonable accuracy, a wide range of modification devices; has a good command of idiomatic expressions and colloquialisms with awareness of connotative level of meaning; backtrack and restructure around a difficulty so smoothly the interlocutor is hardly aware of it.* A further descriptor calibrated in Year 2 continues this emphasis on a precise, fluent and comprehensive command of meaning: *shows great flexibility reformulating ideas in differing linguistic forms to give emphasis, to differentiate and to eliminate ambiguity.*

Broad or Narrow Levels

The finer levels marked out on the difficulty scale as intervals of approximately 1 logit (0.97 in the middle; 1.10 at the ends) were not adopted for the Common European Framework (Council of Europe 1996) which instead uses as Common Reference Levels the "natural levels" introduced earlier.

The finer levels from the study are, however, referred to in the text and indicated in the illustrative scales of descriptors contained in the appendix to the Framework.

The main argument given for adopting fewer, broader bands is often presented as a psychometric one (fewer steps = higher reliability) and the argument for more, smaller defined steps primarily an educational one (more steps = visible progress = motivation) (Hargreaves 1992; North 1992a). However, in discussing scales of language proficiency, it is important to distinguish between the number of decisions any one person is making in a test and the number of levels which exist in the framework as a whole. This important distinction unfortunately tends to get missed in discussion of this issue.

The IELTS test has been criticised (e.g. Hamp-Lyons and Henning 1991) for having 9 levels, which is known to be more than necessary, and more than the raters can handle. Precisely because of this problem, CASE reduced from the IELTS 9 bands to 6 bands (Milanovic et al 1992/6). In Eurocentres, on the other hand, whilst we have a 10 level system, these 10 levels are regrouped for some purposes into 5 categories: Beginner (1), Elementary (2&3), Intermediate (4&5), Upper Intermediate (6&7), Advanced (8–10). For progress and exit assessment we use all the levels, even adding "plus levels" to give 20 rating steps in the system as a whole. However, this is feasible because, since students are in classes by level, it is extremely rare to come across a range of more than three of the ten performance levels in a class even at the end of a course of three months. In practice, even using "plus levels" raters do not have to decide between more than 5 or 6 points (e.g. 3; 3+; 4; 4+; 5) because there will only be a limited range of level in the class. The first step in the assessment procedure used is in fact to identify what that range of level in the class is, in order to concentrate on the right part of the scale.

Whilst it may be true that by using fewer categories it is more difficult to put a person, an examination, or a course in the wrong place and that by using fewer categories one should therefore gain more reliable information, it is also true that such an approach leads to a situation in which standards that may represent a substantial difference in terms of learning hours can be presented as "the same." In other words, in using fewer level categories, accuracy is sacrificed for reliability. A branching approach based on a descriptor bank could circumvent this problem. Firstly, as demonstrated above, the number of narrower levels proposed is itself based upon the reli-

ability of the descriptor scale itself. Those sectors or institutions who prefer to keep to broader levels could do so, but the "reliability" argument against 9 or 10 narrower levels appears flawed. Secondly, with a calibrated descriptor bank, users could in any case "cut" the vertical dimension of the descriptor hierarchy where they wanted to cut it, and create transparent, relevant descriptions of competence for local, sector-specific level systems, which could all be calibrated back to the common framework in one coherent meta-system.

There is in fact a tendency for distinctions between bands to be made finer once a scale takes a framework role. The development of the FSI scale family is a case in point. The original FSI scale was a simple 0–5 scale. By Wild's (1975) publication of the scale "plus levels" were used between the defined levels. By 1983, the "plus levels" in what was now called the ILR scale were fully defined, giving 11 levels. The 9 band ACTFL Guidelines derived from the FSI scale expanded each of the lower two levels into three, and Level 2 into two. Meredith suggested taking this process a stage further by adding a "+" for an above average performance, and a "–" for a below average one (Meredith 1990). What is relevant to the present discussion is that just adding the plus and minus scores to the ACTFL grades *increased* the accuracy of a multiple regression by 4%. This may not seem that much, but

is another indication that fewer levels does not necessarily lead to greater reliability, let alone accuracy.

A Holistic Scale

The holistic scale for Interaction shown in Figure 7.1, was produced from more global descriptors included in the survey.

Figure 7.1: A Holistic Scale for Interaction

Mastery	Has a good command of idiomatic expressions and colloquialisms with awareness of connotative levels of meaning. Can convey finer shades of meaning precisely by using, with reasonable accuracy, a wide range of modification devices. Can backtrack and restructure around a difficulty so smoothly the interlocutor is hardly aware of it.
Effective Operational Proficiency	Can express him/herself fluently and spontaneously, almost effortlessly. Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions. There is little obvious searching for expressions or avoidance strategies; only a conceptually difficult subject can hinder a natural, smooth flow of language.
Vantage Plus	Can use the language fluently, accurately and effectively on a wide range of general, academic, vocational or leisure topics, marking clearly the relationships between ideas. Can communicate spontaneously with good grammatical control without much sign of having to restrict what he/she wants to say, adopting a level of formality appropriate to the circumstances.

Figure 7.1: A Holistic Scale for Interaction (cont.)

Vantage	Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without imposing strain on either party. Can highlight the personal significance of events and experiences and account for and sustain views clearly by providing relevant explanations and arguments.

The Development of a Common Framework Scale of Language Proficiency

Threshold Plus	Can communicate with some confidence on familiar routine and non-routine matters related to his/her interests and professional field. Can exchange, check and confirm information, deal with less routine situations and explain why something is a problem. Can express thoughts on more abstract, cultural topics such as films, books, music etc.
Threshold	Can exploit a wide range of simple language to deal with most situations likely to arise whilst travelling. Can enter unprepared into conversation on familiar topics, express personal opinions and exchange information on topics that are familiar, of personal interest or pertinent to everyday life (e.g. family, hobbies, work, travel and current events).
Waystage Plus	Can interact with reasonable ease in structured situations and short conversations, provided the other person helps if necessary. Can manage simple, routine exchanges without undue effort; can ask and answer questions and exchange ideas and information on familiar topics in predictable everyday situations.
Waystage	Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar matters to do with work and free time. Can handle very short social exchanges but is rarely able to understand enough to keep conversation going of his/her own accord.

Break-through	Can interact in a simple way but communication is totally dependent on repetition at a slower rate of speech, rephrasing and repair. Can ask and answer simple questions, initiate and respond to simple statements in areas of immediate need or on very familiar topics.
Tourist	Can ask and tell the date and time of day, follow short, simple directions and make simple purchases where pointing or other gesture can support the verbal reference.

The descriptors concerned all demonstrated a good degree of model fit and stability across different contexts. They were "very good items" in the quality hierarchy set out at the end of Chapter 6. Some writers would not accept the validity of subsuming individual descriptors into a holistic scale, but as was pointed out at the end of Chapter 2, there are two sides to this argument. If a holistic scale is an overview to a more differentiated system, rather than a substitute for it, it meets a pragmatic need users have to get an idea of what is being described. This is not necessarily an argument for a single global proficiency scale; holistic scales for Reception, Interaction and Production can serve the same purpose.

Scale Shrinkage

As discussed in Chapter 3, classical measurement recognises four types of scales: nominal (for category data); ordinal (ranking); equal interval (with mathematically constant units), and ratio (an equal interval scale starting at absolute zero). A fifth type, linear, *could* be added if one considers that a scale (e.g. a standardised scale) can be linear without necessarily having equal intervals, or that a defined scale could have steps of unequal intervals which are in a defined linear relationship to one another.

The logit scale produced in the analysis should, according to Rasch measurement theory, be able to be interpreted as a virtually equal-interval scale of proficiency, though as mentioned, Hambleton et al (1991: 87) add the rider that although it is "popular and reasonable" to assume this, this is nevertheless not strictly true. However, two notes of caution must be added before the linearity of the scale is interpreted too literally. Firstly, a number of distortions in the measurement scale have been identified and, hopefully, corrected. The coherence of the result suggests that those corrections succeeded.

Secondly, however, there is a kind of distortion known as "scale shrinkage" which has not yet been discussed. Yen (1985: 403) has demonstrated with simulated and real test material that when (psychometric) unidimensionality is assumed in an analysis of what is actually (psychologically) multidimensional data, the effect is to cause a shrinkage of the scale as the level of proficiency increases. Considering that the scale is on the very borderline of psychometric unidimensionality, it is quite likely that the scale does just this. Taking "*Smattering*," the point below *Tourist*, for which there was only one descriptor *can use some basic greetings; can say yes, no, excuse me, please, thank you, sorry (-5.68)* as a zero starting point, it appears somewhat surprising that *Threshold* is the 5th unit on the scale, and in fact half way to a level identified as *Mastery*. Eurocentres experience with its own scale, and the relationship between that scale, public examinations and Council of Europe specifications would suggest that *Threshold* (ALTE Level 2) is more likely to be half way towards at best what is provisionally labelled *Effective Operational Proficiency*, intended to correspond to the level of the Cambridge Certificate of Advanced English (ALTE Level 4).

An alternative explanation might be that the scale is more distorted at the bottom (i.e. with minus logit values) than it is at the top (with positive logit values). When we come to discuss calibrating learners, we will see that there are at least grounds for suspecting that this might be the case with this

scale. Warm (1989: 427–8) discusses this point, stating that Lord (1983) found that when the logit difficulty/ability scale distorts towards the ends (as found in this study) it does so in an asymmetric fashion so that the negative values are more distorted. If such a distorted scale is treated as if it were a linear, equal interval scale, or if non linearity is adjusted for in a symmetrical fashion, as was attempted in setting cut-off points between levels (0.97 logits interval in the middle; 1.10 logits at the ends) then it might be that such symmetrical treatment underestimates the size of the upper half of the scale, creating the "scale shrinkage" to which Yen (1985) refers.

These three arguments for possible scale shrinkage towards the top of the scale (*theoretical*: hidden multidimensionality (Yen); *technical*: asymmetrical scale bias (Warm); *pragmatic*: comparison to other sets of levels and to progress norms suggest that the decision by the Council of Europe Framework Working Party to adopt the levels described above as *Effective Operational Proficiency* and *Mastery* as full levels when reducing the number of levels for the Common Reference Levels may actually *increase* the linearity of the scale of levels rather than reducing it. The issue is not an entirely straightforward one.

If the logit scale *is*, however, actually weighted slightly towards the bottom and shrunken at the top, for any of the reasons suggested, this could be seen as having its advantages. Some teachers in the workshops were really quite sceptical about the possibility of providing descriptors of communicative language proficiency for learners who had studied for only 80 or 160 hours, and people seem to generally find it easier to describe what people with a high level of proficiency can do, and what people at a modest level of proficiency *cannot* do. Secondly, it could be argued that a Swiss and/or European Framework would not be damaged if, whilst being comprehensive enough to encompass high level learning, it had a slight weighting towards the lower half of the scale, where the vast majority of the learners are to be found. In such a way, more milestones would be able to be provided at lower levels in order for learners to see progress.

It could also be the case that what is being illustrated here is a fairly equal interval vertical dimension which is not shrunken towards the top at all, but which does not take adequate account of the greater breadth of language at higher levels. Both Trim (1978: 25) and Lowe (1985: 21–22) present progress in language proficiency with diagrams like the cone formed by an ice cream cornet, and perhaps the linearity should be interpreted in the

sense of the vertical axis of the cone rather than strictly in terms of learning load, seat time.

Content Coherence

As can be seen from the arguments in the last section, producing a scale purely from statistical data in the search for equal intervals (the first method of establishing cut-offs) is not actually guaranteed to lead to a true equal interval scale because of the known distortions in the logit scale and the possibility of shrinkage towards the top. Therefore, as was the case during the analysis, the statistical information available has to be interpreted with informed judgement. An obvious way to do this was to check the coherence of the content of the different levels.

The main way in which the coherence was checked was by studying the descriptors for a particular category (e.g. Listening in Interaction) in order to see what elements the descriptors seemed to have largely in common, to then pick the descriptors apart into these elements and then to list the elements in tables. These charts enabled a visual check on (a) the logic of progression in each category and (b) the consistency with which an elements like for example “everyday situations of a concrete type” appeared in the same band on tables for different descriptive categories. For example, the items in the sub-scale for Listening in Interaction, listed in Table 7.3 were

split into the elements "Setting," "Speech" and "Help" as shown in Table 7.4.

A very clear progression is visible in all three columns. Speech must at first be *very clear, slow, carefully articulated repeated speech directed at* the recipient. Then comes *clear, slow, standard speech, directed at him/her* followed by *clearly articulated standard speech* (which no longer has to be especially slowly or carefully adjusted for him/her) and finally the *standard spoken language*.

Table 7.3: Listening in Interaction: Sub-scale of Descriptors

Level	Logit	Descriptor	Source
V+	2.56	Can keep up with an animated conversation between native speakers.	North 7
V	1.11	Can understand in detail what is said to him/her in the standard spoken language even in a noisy environment.	North 6 / Hoff V / Lon5/FSI 3
T+	0.33	Can extrapolate the meaning of occasional unknown words from the context and deduce sentence meaning provided the topic discussed is familiar.	ASLPR1 +
T	-1.04	Can generally follow the main points of extended discussion around him /her, provided speech is clearly articulated in standard dialect.	North4 / AMES3 / elviri3
T	-1.09	Can follow clearly articulated speech directed at him/her in everyday conversation, though will sometimes have to ask for repetition of particular words and phrases.	IELTS5 edited
W+	-1.83	Can understand enough to manage simple, routine exchanges without undue effort.	AMES3
W+	-2.13	Can generally identify the topic of discussion around her which is conducted slowly and clearly.	North3
W+	-2.13	Can generally understand clear, standard speech on familiar matters directed at him, provided he/she can ask for repetition or reformulation from time to time.	wilk3/ES U 4 / EC5 / llb1 edit /North4

W	-2.72	Can understand what is said clearly, slowly and directly to him/her in simple everyday conversation; can be made to understand, if the speaker can take the trouble.	EC3 / finn3 / HoffII

Table 7.3 (continued): Listening in Interaction: Sub-scale

Level	Logit	Descriptor	Source
B	-3.5	Can understand everyday expressions aimed at the satisfaction of simple needs of a concrete type, delivered directly to him or her in clear, slow and repeated speech by a sympathetic speaker.	elviri1/ES U1
B	-3.64	Can follow speech which is very slow and carefully articulated, with long pauses for him/her to assimilate meaning.	New

B	-4.12	Can understand questions and instructions addressed carefully and slowly to him/her and follow short, simple directions.	EC1

Table 7.4: Listening in Interaction: Calibrated Elements

Lvl	Setting	Speech	Help
V+	-animated conversation between native speakers		
V	-even noisy environments	-standard spoken language	
T+	(topics which are familiar)	*As Threshold*	-none; extrapolate unknown words; deduce meaning
T	-extended everyday conversation	-clearly articulated standard speech	*As Waystage +*
W+	-simple, routine exchanges -familiar matters	*As Waystage*	-ask for repetition & reformulation
W	-simple everyday conversation	-clear, slow, standard, directed at him	-if partner will take the trouble

Table 7.4: Listening in Interaction: Calibrated Elements (cont.)

Lvl	Setting	Speech	Help
B	-everyday expressions aimed at the satisfaction of needs of a concrete type -short, simple questions & instructions	-very clear, slow, carefully articulated repeated speech directed at him	-sympathetic partner -long pauses to assimilate meaning

A total of 17 such charts were produced and studied for discrepancies.

Where categories appeared to have related elements they were juxtaposed on the same page to aid comparison and check consistency. The coherence shown by the different charts was extremely high, with one single contradiction. This was identified when Grammatical Accuracy and Monitoring & Repair Strategies were juxtaposed on a chart labelled Language Awareness, shown in Table 7.5.

Table 7.5: Language Awareness: Calibrated Elements

Lvl	Controlled Range	Type of Error	Monitoring and Repair
M			-backtrack and restructure around a difficulty so smoothly the interlocutor is hardly aware of it
EOP			

V+	-good grammatical control	-occasional “slips” or non-systematic errors and minor flaws in sentence structure occur rarely	-often corrected in retrospect

Table 7.5 (cont): Language Awareness: Calibrated Elements

Lvl	Controlled Range	Type of Error	Monitoring and Repair
V		-slips and errors (which he can generally correct)	-if he becomes conscious of them -if it is a “favourite mistake” consciously monitored for -if it has led to misunderstanding
T+	-reasonable accuracy in familiar contexts; generally good control	-no mistakes which lead to misunderstanding; clear what trying to express -noticeable mother tongue influences	
T	-reasonably accurate use of a repertoire of frequently used “routines” and patterns associated with more predictable situations		
W+			

W	-some simple structures used correctly -limited control of even short, simple sentence structures	-systematically makes basic mistakes -tends to mix up tenses and forget to mark agreement; nevertheless it is clear what he/she is trying to say

At *Threshold Plus* one is said *not to make mistakes which lead to misunderstanding*, while at the next level, *Vantage*, one is said to be able to *correct mistakes which have lead to misunderstanding*—which should not have happened, should they? This may be an inconsistency in the scaling, or it may rather reflect the fact that the linking of mistakes to "misunderstandings" is in any case a slightly questionable concept. The inevitable mistakes at a low level are possibly less likely to lead to misunderstandings than the occasional bad choice of phrase or wrong tense from an advanced learner whose language is more likely to be accepted at face value. Ironically that descriptor No (226) is one of the few items formulated from a statement made during the teachers' discussion of aspects of learner performance in the workshops.

The content of all the charts showed a systematic progression up each column through the different levels, as discussed for Listening in Inter-

action. With the sole exception mentioned, the consistency between charts is very striking. This degree of coherence strongly suggests that the corrective measures taken to compensate for scale distortion (see Chapter 5) were in fact successful, and that the cut off points established between levels were sensible.

Progression in Proficiency

Now that a set of levels at more or less equal intervals had been established on the scale, and after the confirmation that the descriptor content of those levels was coherently organised both vertically in terms of progression and horizontally in terms of relationships between categories, one was in a position to see what the scale had to say about the nature of developing language proficiency as perceived by teachers of English in Switzerland.

Of course one has to be careful to make any statements relative. What has been achieved by the pre-testing and refinement of descriptors in workshops and the calibration of descriptors with satisfactory properties with a measurement model has been to objectively scale an inter-subjective consensus. The picture offered by the scaled descriptors cannot be accepted as the full picture for a number of reasons.

Firstly, the descriptors mainly reflect the style of what Bachman (1990a: 303–14) has called the "real life" approach to language assessment. Satisfactory statistical properties of descriptors and thus successful calibration means that the raters were able to use this style of description to rate their learners. Whilst this offers evidence of the validity of this approach for this purpose in this context, it does not necessarily mean that the picture of proficiency offered is "true."

Secondly, the original descriptors from which the descriptors used in the survey were edited came themselves from particular sources and reflect the house styles of those sources. Styles which proved counter-intuitive to the 100 teachers involved were excluded in the workshop pre-testing or during the analysis. One third of the surviving 212 items come purely from the current writer and Eurocentres, two thirds come either purely or partly from the current writer and Eurocentres. This suggests that the Eurocentres house style of short, positively worded descriptors "works" for Swiss teachers, which gives it a certain pragmatic validity. This does not necessarily mean that what is being described is "true" but rather that it reflects to a satisfactory degree the way in which teachers think about the proficiency of their learners.

Thirdly, the coverage of the calibrated descriptors shows noticeable gaps. Some of these gaps may be purely accidental—descriptors are just lacking, whilst others may be significant—there being nothing to describe at that point. The fact that absences may or may not be significant means one needs to be cautious in generalisations about proficiency.

Finally and fundamentally, whilst the picture presented is certainly interesting, there is no guarantee that the results offer a picture of developing second language acquisition. Ingram (1985: 221–30) has been criticised, particularly by Pienemann and Johnson (1987: 91–97) for apparently claiming that a proficiency scale like the ASLPR could offer such a picture. One needs to bear in mind that as mentioned above all the descriptors come from somewhere and were not produced though an detailed analysis of the discourse of the target population as were, for example, those of Fulcher (1993). In addition, as mentioned above the descriptors could be claimed to reflect the way teachers relate to proficiency, thus saying as much about the way the teachers think as about the proficiency being assessed at second hand, (c) the results are the product of a cross-sectional analysis at one point in time, whereas, as (Larsen-Freeman and Long (1991: 267) admit in relation to criticisms of the 1970s morpheme studies (Dulay and Burt 1974; Bailey et al 1974; Hakuta 1974/8; Larsen-Freeman 1978), it is necessary to conduct longitudinal studies of the progress of individual learners to gain

insights into the nature of the SLA process. Finally, as Nunan (1989: 85) points out in criticising the claims of Adams et al (1987) to have produced a snapshot of SLA development through a Rasch analysis of interview tasks, one needs to be cautious in making claims about developmental orders based upon the fit of data to a statistical model.

Given all these caveats, the results are still very interesting, and remarkably coherent.

Holistic Overviews

One of the 17 charts produced by breaking up the descriptors into constituent elements was a collation of the more generalisable content from the other 16. That is to say, in a final check of coherence between what were by this stage being thought of as sub-scales for the various categories, the content of all the 16 charts produced directly from the sub-scales was carried over and summarised onto one "global" chart shown as Table 7.6.

To illustrate the way the global chart was arrived at, one can consider the column Topic/Settings. The statement at the top, at the level labelled V+ (*Vantage Plus*) *a wide range of general, academic, vocational or leisure topics*, was already on the chart headed "global." This is because it comes from a definition, No 241, *can use the language fluently, accurately and effectively on a wide range of general, academic, vocational or leisure topics*, which is on the sub-scale of more holistic statements headed "Overall Interaction." in Appendix 3. The entry at the level underneath, *most general topics*, originates from No 204 on the sub-scale for Conversation: *can engage in extended conversation in a clearly participatory fashion on most general topics.* The next entry, at T+, *familiar matters within his/her field* comes from the chart for "Transactions" and descriptor No 231 on the sub-scale for Information Exchange: *can exchange accumulated*

factual information on familiar routine and non-routine matters within his field with some confidence.

Table 7.6: Global Calibrated Elements

Lvl	Action	Topic / Setting	Limitation
M	-good command of idiomatic expressions and colloquialisms with awareness of connotative level of meaning -convey finer shades of meaning precisely		
EOP	-express him/herself fluently and spontaneously, almost effortlessly -produce clear, smoothly flowing well-structured speech		

Table 7.6 (continued): Global Calibrated Elements

Lvl	Action	Topic / Setting	Limitation
V+	-communicate spontaneously, often showing remarkable fluency and ease of expression -adopt a level of formality appropriate to the circumstances -use the language fluently, accurately and effectively	-a wide range of general, academic, vocational or leisure topics	
V	-interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without imposing strain on either party	-most general topics	-can be hesitant as he or she searches for patterns and expressions

T+	-exchange, check and confirm information with confidence -deal with difficult, less routine situations and explain why something is a problem -summarise, and express thoughts and reactions to films, music, articles, short stories etc.	-familiar matters within his/her field	-may need repetition if response rapid or extended -limited precision -generally good control, though with noticeable mother tongue influences
T	-flexible exploitation of a wide range of simple language -maintain conversation & discussion -deal with most situations likely to arise when making travel arrangements through an agent or when actually travelling	-topics that are familiar or of personal interest -most topics pertinent to his everyday life	-pausing for grammatical and lexical planning and repair is very evident -needs to ask for repetition and reformulation

Table 7.6 (continued): Global Calibrated Elements

Lvl	Action	Topic / Setting	Limitation
T cont.		(family, hobbies and interests, work, travel, and current events)	-expresses main point comprehensibly but may sometimes be difficult to follow when trying to say exactly what he /she would like to
W+	-participate in short conversations -make him/herself understood and exchange ideas and information	-common aspects of everyday living -on familiar topics in predictable everyday situations	-if he/she can ask for help to express what he wants to, otherwise will generally have to compromise the message -if can ask for repetition and reformulation
W	-handle very short social exchanges -get simple information about travel, use public transport -exchange limited information -use simple phrases and get what he needs in common, simple everyday situations	-simple, routine, direct exchange of information -simple needs of a concrete type -basic commun-icative needs	-rarely able to keep a conversation going, but can be made to understand if the partner will take the trouble -given help
B	-ask and answer simple questions - initiate and respond to simple statements -interact in a simple way	-needs of a concrete type -areas of immediate need -very familiar topics	-communication is totally dependent on very clear, carefully articulated repetition of simple language at a slower rate of speech

Table 7.6 (continued): Global Calibrated Elements

Development of a Common Framework Scale of Language Proficiency

Lvl	Action	Topic / Setting	Limitation
B Cont Tour	-make simple purchases, supported by gesture -ask and tell day, time of day and date		rephrasing and repair

As regards the two entries for *Threshold*, the first: *topics that are familiar or of personal interest* comes again from the "Conversation" chart: No 166: *can initiate, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest* and from No 152: *Can enter unprepared into conversations on familiar topics.* The second element, *most topics pertinent to his everyday life such as family, hobbies and interests, work, travel, and current events,* comes from the chart for "Range" in the column "Setting" and originates from No 225 on the sub-scale for Vocabulary Range: *has a sufficient vocabulary to express him/herself with some circumlocutions on most topics pertinent to his everyday life such as family, hobbies and interests, work, travel, and current events.*

For Waystage Plus entries come on the one hand from the "global" the sub-scale for Overall Interaction, and on the other from Service Encounters. With the number of comments about topics in descriptors at Waystage and Waystage Plus, entries on this global overview chart are more and more selective. However, the level at which identical or similar content elements from different sub-scales are placed appears totally coherent. This "global" chart thus offers a holistic overview of proficiency in interaction at the different levels. and was used as a guide in drawing up the holistic scale for interaction.

To recap, the topics/settings column shows the same kind of clear progression commented on in relation to Listening Comprehension in Interaction. At *Tourist*, the learner does not yet possess a generalisable "level" of language sufficient to cope with topics, as opposed to isolated tasks. At *Breakthrough* the learner has sufficient language to cope with *immediate needs of a concrete type* and *very familiar topics*. At *Waystage*, this extends to *basic communicative needs* and to the *simple, routine, direct transfer of information*. At *Waystage Plus basic communicative needs* has broadened to *common aspects of everyday living*, and topics are now described as *familiar* (as opposed to *very familiar*), *topics in predictable everyday situations*. By *Threshold* the topics are *familiar (pertinent to his everyday life* or *of personal interest)*, the latter being a bridge to the capacity at *Threshold Plus* (*Independence*) to deal with familiar matters within his/her field. In other words, the learner can now deal with a range of topics which happen to be of personal relevance. This could also be said to reflect the tendency in many scales (including Eurocentres) to start making comments at this level about being able to communicate in professional life. At *Vantage* the learner can now deal with *most general topics* and at *Vantage Plus*, in the last comment on topics, the learner can handle *a wide range of general, academic, vocational or leisure topics*.

This progression could be represented schematically as in Table 7.7.

Table 7.7: Topics and Settings: Calibrated Elements

Level	Elements
M	*No restriction*
EOP	*No restriction*
V+	-wide range of general, academic, vocational topics
V	-most general topics

T+	-familiar; within (professional) field of interests
T	-familiar everyday life; -topics of personal interest
W+	-common predictable everyday; -familiar
W	-routine information exchange; -basic communicative needs
B	-concrete immediate; -very familiar

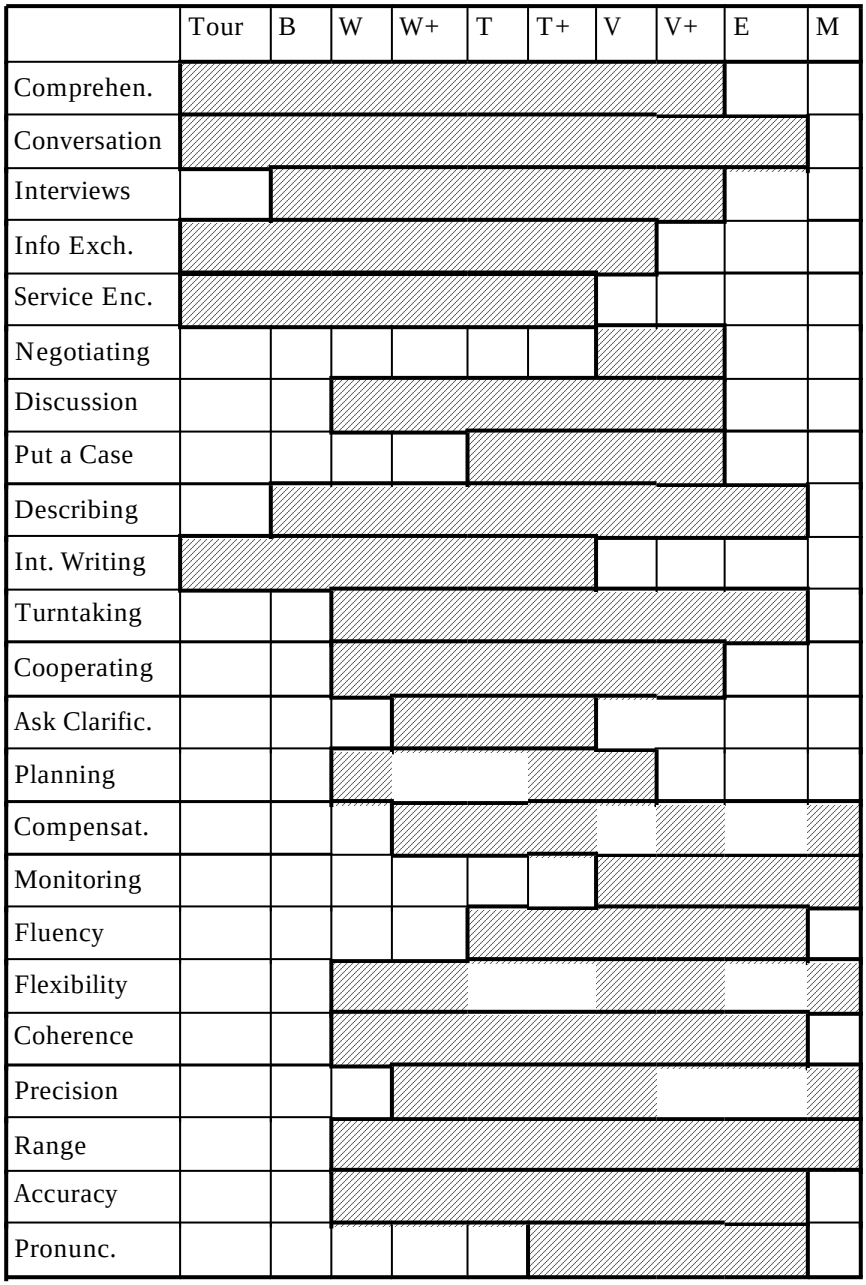
or in an even more abstract fashion, just to demonstrate at which levels the issue of "topic/setting" is defined, as in Figure 7.2.

Figure 7.2: Coverage of Topics / Settings

Topics, Settings	Tour	B	W	W+	T	T+	V	V+	EOP	M

Such an abstract overview chart showing the levels at which relevant descriptors are situated is given for the 23 principal content areas in Figure 7.3.

Figure 7.3: Coverage of Categories of Descriptors



It is of course very difficult to draw conclusions from the absence of a particular area at a particular level. There would seem to be at least four different explanations why a gap may occur.

The feature:

a. exists at this level: some descriptors were included in the survey, but were dropped in quality control
b. probably exists at this level: descriptors could presumably be written, but have not been
c. may exist at this level: but formulation seems to be very difficult if not impossible
d. does not exist or is not relevant at this level

Even with these provisos about interpreting gaps, the chart does present an interesting picture of the nature of proficiency at different levels.

It is striking that descriptors for *Tourist* and *Breakthrough,* despite a very narrow range, do manage to cover both transactional and interpersonal language use. However, it is equally striking that there are no qualitative descriptors at all. All descriptors for strategies and analytic aspects of proficiency start at *Waystage*. Conversely, the top two levels, *Effective Operational Proficiency* and *Mastery* are covered by very few descriptors. Those descriptors which do exist tend to be focused on Discussion and aspects of coherent language production on the one hand, and on analytic aspects of proficiency and self repair on the other. It is pretty obvious that Negotiating belongs to this group as well.

The vast majority of descriptors have been calibrated in the range between *Waystage* and *Vantage Plus*. That is to say most of the descriptors have been calibrated in the range covered by the three Council of Europe specifications *Waystage*, *Threshold* and *Vantage*. The amount of detail described in *Waystage*, *Waystage Plus* and *Threshold* (accounting for 122 of the 212 descriptors) is not really surprising since *Waystage* and *Threshold* have been around for 20 years and have been very influential in the development of many of the scales used as sources.

The level at which each content area is first described also seems to show a logical pattern. Expressing opinions in discussion, Turntaking and Cooperating all start at *Waystage*, which seems reasonable. However, it is rather surprising that Asking for Clarification should first appear at *Waystage Plus*. In fact only one descriptor in this area was dropped (*Can ask someone to give more information*) which does not appear to be noticeably easier than the two descriptors at *Waystage Plus*: *Can ask very simply for repetition when he or she does not understand* and *Can ask for clarification about key words not understood using stock phrases.* That Putting a Case is said to start at *Threshold* and Negotiating to start at *Vantage* seems unsurprising, as does the fact that below *Waystage Plus* there is little to say about Precision. The suggestion that successful Monitoring & Repair starts above *Threshold* is discouraging, but not illogical.

Categories of Language Use

Each content area shows an apparently systematic progression, like that outlined for Topics / Settings earlier. For example, the straightforward progression shown in Table 7.8 for Information Exchange reminds one of the concept of concentric spheres of involvement associated with audio-visual methods. The only surprise is *Threshold* where once might expect a little more than just directions.

Table 7.8: Information Exchange: Calibrated Elements

Level	Elements
T+	-accumulated factual info on familiar matters within field
	-describe how to do something, giving detailed instructions
T	-detailed directions
W+	-simple directions & instructions
	-pastimes habits, routines
	- past activities

W	-simple, routine direct
	-limited work & free time
B	-themselves & others
	-home
	-time
Tour	-day, time of day, date

Table 7.9 for Service Encounters shows a certain amount of overlap between levels, especially between *Breakthrough* and *Waystage*. There seem to be three phases: *Basic User* (*Tourist-Waystage Plus*) concerned with getting around shops and transport; *Threshold* (including *Threshold Plus*) dealing with less common, less routine and potentially situations, and then *Vantage* (including *Vantage Plus*), the surviving Negotiating items, which deal with resolving problems and disputes.

Table 7.9: Service Encounters: Calibrated Elements

Level	Elements
V+	-negotiate solution to dispute
	-outline case for compensation
V	-explain problem and make clear expect a concession

T+	-deal with less routine situations in shops
	-return unsatisfactory purchase
T	-make a complaint
	-deal with likely though less routine travel situations
W+	-get straightforward info
	-survival & routine travel needs
W	- simple transactions / purchases / enquiries
	- get simple info
	-quantities, numbers, prices
B	-ask for / give things
	-quantities, numbers, prices
Tour	-simple purchases supported by gesture

The scale for Describing and Narrating summarised in Table 7.10 was discussed in Chapter 6 because of the fact that some of the descriptors were calibrated considerably lower than the level at which they had been placed in the source scale, the ASLPR. The progression of topics is, however, very reminiscent of that for Information Exchanges. The gap at levels *Threshold Plus* and *Vantage* is caused by the fact that many of the descriptors were calibrated lower than expected. The following adapted Eurocentres descriptor was successfully calibrated at *Vantage* in 1995: *Can give clear, detailed descriptions on a wide range of subjects related to his/her field of interest.*

The sub-scale for Conversation shown in Table 7.11 is somewhat more complicated. It is divided into what appear to be the main elements in the descriptors: Setting; Topic/Register; Specific microfunctions (illocutionary), and Limitations. The range of settings progresses from *very short social exchanges* through *simple, restricted, face-to-face conversation* and *participating in short conversation* to *simple face-to-face conversation, the ability to enter into conversation unprepared* and *maintain conversation and discussion (*at *Threshold).* Then there is a gap, possibly a plateau, until at *Vantage* the learner is said to be able to *sustain relationships with native speakers without unintentionally irritating or amusing them*, and to *engage in extended conversation in clearly participatory fashion*, then another gap/plateau until at *Effective Operational Proficiency*, the learner is said to be able to *use the language flexibly and effectively for social purposes —including emotional, allusive, joking usage.*

Table 7.10: Describing and Narrating: Calibrated Elements

Level	Elements

The Development of a Common Framework Scale of Language Proficiency

EOP	-clear, detailed description of complex subjects
V+	
V	- clear, detailed descriptions on a wide range of subjects related to his/her field of interest
T+	-basic details of unpredictable occurrences e.g. accident
T	-on a variety of familiar subjects within his/her field of interest -plot of book/film -experiences -reactions to both -dreams, hopes, ambitions -a story
W+	-objects, pets possessions -events & activities -likes / dislikes -plans / arrangements -habits / routines -personal experience
W	-people, appearance, background, job -places & living conditions
B	-where they live
Tour	

Both *Waystage* and *Threshold* have a proviso added. In the former case this is difficulty in keeping the conversation going (not yet quite able to *initiate, maintain and close simple, restricted face-to-face conversation*) and in the latter case, at *Threshold, sometimes being difficult to follow when trying to say exactly what he/she wants to.* This makes sense because the learner is not yet quite able to *express his/her thoughts about abstract subjects e.g. music, films.*

A number of microfunctions appear in descriptors classified under "Conversation" either as a descriptor in their own right or as part of a more holistic descriptor, and in view of the often mentioned difficulty in deciding an order of progression amongst illocutionary functions, the progression deduced from the teacher judgements of difficulty is very interesting. Where a function appeared as a descriptor on its own, teachers pointed out that it could of course be in speech or writing, but that they tended to associate them with speech unless stated otherwise. In the questionnaires it was made clear speech was meant since they were included in the list of "Spoken Tasks."

Table 7.11: Conversation: Calibrated Elements

Level	Elements
EOP	-emotional, allusive, joking usage
V+	
V	-convey degrees of emotion
	-highlight personal significance of events
T+	-express thoughts about abstract subjects, e.g. music, films
T	-express e.g. surprise, happiness, sadness, interest, indifference
W+	-ask for give or refuse permission
W	-express how he/she feels
	-offers, invitations
	-apologies thanks, polite greeting, farewells, intros
B	-make introductions
	-basic greeting and leave-taking
Tour	-some basic greetings

Apart from showing a credible progression, the functions seem to fall into four bands: conventional social formulae (*Tourist and Breakthrough*); expression of everyday social illocutionary functions, as found in all textbooks (*Waystage & Waystage Plus*); personal expression: feelings and thoughts (*Threshold and Threshold Plus*); differentiated personal expression: (*Vantage*);

effective, flexible use of language for social purposes (*Effective Operational Proficiency*).

Turntaking Strategies are shown in Table 7.12. Up until *Threshold*, the focus is on getting in on the act. At *Threshold* the learner can turntake effectively in simple conversation. At *Vantage*, the point where personal expression becomes more differentiated, the learner can *take an active part in informal discussion*, can *account for and sustain his/her views*, and *sustain relationships with native speakers without unintentionally irritating or amusing them*, and so *engage in extended conversation in clearly participatory fashion*, Turntaking becomes more independent, more strategic *(gain time to keep the floor)*, the aspect that comes to the fore by *Effective Operational Proficiency*.

As can be seen from Table 7.12, effective Cooperation Strategies begin at *Threshold* when the learner is able to *maintain conversation and discussion* and when he/she *can express* (or repeat) *the main point he/she* (or someone else) *is making comprehensibly*.

Table 7.12: Turntaking Strategies: Calibrated Elements

Level	Elements
EOP V+	-select from fluent repertoire to get floor, gain time, keep floor

V	-gain time to keep turn -take turn/intervene when appropriate -initiate / end when needs to (not always elegantly)
T+	
T	-initiate, maintain, close simple face-to-face
W+	-simple techniques to start, maintain, end
W	-ask for attention

Judging from the very, very slight difference between the two descriptors for *Threshold* and *Vantage*, there then appears to be a plateau when the learner can converse or work in group discussion. This is followed at *Vantage Plus* by the ability to follow up and refer back skilfully to others' contributions, weaving one's own offering into the joint discourse.

Table 7.13: Co-operating Strategies: Calibrated Elements

Level	Elements
EOP	
V+	-relate own contributions skilfully to others' -follow up statements & inferences -give feedback
V	-help discussion along on familiar ground: confirming comprehension, inviting other in etc.
T+	
T	-help keep discussion on course by repeating back -invite into discussion
W+	
W	-indicate when following

Compensatory Strategies appear to fall into three bands: Firstly appeal to paralinguistic aid. Secondly, at around *Threshold*, interactive strategies: trying something and getting fine-tuning from the interlocutor. All three of Kellerman et al's (1987) types are present: Linguistic (foreignise); Approximative (word with similar meaning) with the finesse of qualifying it (e.g. a bus = "truck for people") at the higher level, and Analytic: defining attributes. Finally, in the third phase, overt substitution is no longer necessary. Resources are sufficient to permit circumlocution, so skilled at the higher level (*Mastery*) that the interlocutor is hardly aware of it.

Table 7.14: Compensating Strategies: Calibrated Elements

Level	Elements
M	-restructure round difficulty so smoothly interlocutor hardly aware of it
EOP	-gaps readily overcome with circumlocution
V+	-use circumlocution and paraphrase to cover gaps in vocabulary and structure
V	
T+	-qualify a word with similar meaning -define features of something concrete
T	-simple similar word and invite "correction" -foreignise and ask for confirmation
W+	-inadequate word plus gesture
W	

It is a little surprising and disappointing that learners at *Waystage* do not appear to be considered capable of communication strategies; but the progression is logical. The lack of entries for Waystage in the treatment of compensating strategies in existing scales was bemoaned when putting together the descriptor pool, but may not be so far wrong after all.

The fact there appears to be a gap at *Vantage* before the third, circumlocution phase is probably not significant. Substituting alternative means of expression to avoid breakdown is of course directly related to Fluency. Descriptors for Fluency at this level suggest that learners are capable of steering round difficulties too, as shown in Table 7.15. The lowest level Fluency descriptor: *can manage comprehensible phrases with some effort, false starts and repetition* was excluded as it was an unstable anchor item. As commented in Chapter 6, it is not easy to see why. The lowest calibrated descriptor in Table 7.15 is thus at the point (*Threshold*) where the learner can keep going. The next development is at *Vantage* (*fairly even tempo*) with *spontaneous* being the common factor between the two higher levels. The gap at *Threshold Plus* was plugged in the follow up survey with another descriptor: *Can express him/herself with reasonable ease. Despite some problems with formulation resulting in pauses and "cul-de-sacs" he/she is able to keep going effectively without help.*

Table 7.15: Fluency: Calibrated Elements

Level	Elements
M	
EOP	-natural smooth, fluent, spontaneous, almost effortless
V+	-spontaneous, often showing remarkable fluency and ease of expression
V	-stretches of language with fairly even tempo -no strain on either party
T+	-express him/herself with relative ease. - despite pauses and "cul-de-sacs", able to keep going effectively without help
T	-keep going comprehensibly but pausing evident - make self understood in short contributions, even though pauses, false starts and reformulation are very evident
W+	
W	

Finally, Fluency and Compensating Strategies are related to linguistic range, the three being ways of looking at a related phenomenon from pragmatic, strategic and linguistic perspectives respectively. Table 7.16 shows different elements of Range at the different levels. The first column on the chart echoes the category "topics / settings" presented at the beginning of this section, and the progression in topics which can be described. The De-

scription sub-scale, the sub-scale for Information Exchange and that for Range in fact demonstrate considerable coherence, as shown in Table 7.17.

The similarity for *Breakthrough*, *Waystage* and *Waystage Plus* is considerable: "*themselves and other people*" is put at *Breakthrough* as well as *Waystage* in Information Exchange, as in the other two sub-scales, but this makes sense in connection with simple "ping-pong" question and answer interaction involved.

Table 7.16: Range: Calibrated Elements

Level	Settings	Language	Limitations
M		-a good command of idiomatic expressions and colloquialisms with awareness of connotative level of meaning	-little obvious searching for expressions or avoidance strategies.
EOP		-a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions	

V+			-little sign of having to restrict what he /she wants to say
V T+			
T	-most topics pertinent to his everyday life (family, hobbies and interests, work, travel, and current events)	-enough language to get by -a wide range of simple language	-lexical limitations cause repetition and difficulty with formulation at times -need for some circumlocution -major (lexical) errors when expressing more complex thoughts
W+	-routine, everyday transactions involving familiar situations and topics	-well rehearsed memorised simple phrases -a repertoire of basic language	-will generally have to compromise the message and search for words

Table 7.16 (cont.) : Range: Calibrated Elements

Level	Settings	Language	Limitations
W+ cont	-everyday situations with predictable content		

W	-basic communicative needs / simple survival needs -concrete everyday needs -simple needs of a concrete type: personal details, daily routines, wants and needs, requests for information. -predictable survival situations	-basic sentence patterns -memorised phrases, groups of a few words and single expressions and formulae -a limited / narrow repertoire of short memorised phrases and sentences	-frequent breakdown / misunderstanding in non-routine situations

Learners would not be capable of giving a coherent description, but can pose or answer simple information requests. The *Threshold* descriptors on the three sub-scales have parted company, all are talking about different things, but there is no contradiction. This divergence continues with *Threshold Plus* where what seems to come into play is unpredictability, an element of precision, and quantity of information.

Table 7.17: Range; Description; Information Exchange: Calibrated Elements

Level	Range Settings	Describing	Info Exchange
M			
EOP		-clear detailed description of complex subjects	
V+			

Table 7.17 (cont.): Range; Description; Information Exchange: Calibrated Elements

Level	Range Settings	Describing	Info Exchange
V		- clear, detailed descriptions on a wide range of subjects related to his/her field of interest	
T+		-on a variety of familiar subjects within his/her field of interest -basic details unpredictable occurrences e.g. accident	-accumulated factual info on familiar matters within field -describe how to do something, giving detailed instructions
T	-most topics pertinent to everyday life: family hobbies interests, work travel, current events	-plot of book/film -experiences -reactions to both -dreams, hopes, ambitions -a story	-detailed directions
W+	-routine everyday transactions -familiar situations & topics -everyday situations with predictable content	-objects, pets possessions -events & activities -likes/dislikes -plans/arrangements -habits /routines -personal experience	-simple directions & instructions -pastimes habits, routines - past activities

W	-basic commun. needs -simple / predictable survival -simple concrete needs: pers. details, daily routines, info requests	-people, appearance -background, job -places & living conditions	-simple, routine direct -limited work & free time

Table 7.17 (cont.): Range; Description; Information Exchange: Calibrated Elements

Level	Range Settings	Describing	Info Exchange
B		-where they live	-themselves & others -home -time
Tour			-day, time of day, date

Returning to the main Range sub-scale in Table 7.16, a clear progression is also evident in the columns Language and Limitations. Limita-

tions progress from *frequent breakdown and misunderstandings* to *message compromise* to *repetition, some difficulty in formulation and need for circumlocution* to *little sign of restrictions to little obvious searching or avoidance strategies*. The different in difficulty between the last two seems a little small, accounted for by the fact that the top element is the end of what is a very high level descriptor, but on the other hand even native speakers search for expressions and use avoidance strategies.

The most striking thing about the Range chart is that it jumps from what is defined as a brilliant elementary performance (*wide range of simple language*) to an advanced performance (*good command of a broad repertoire*). "Normal" range does not appear. This problem is being caused by the lack of definitiveness of statements like "wide range," which were not popular with teachers. An attempt was made in 1995 to plug the gap with the following descriptor: *can vary formulation of what he/she wants to say and can use some complex sentence forms*, but this too came out calibrated at *Vantage Plus.* This demonstrates the way in which, even when once one has a framework of coherently calibrated descriptors and one tries to find or formulate new descriptors to plug gaps, this may not always be successful.

The Development of a Common Framework Scale of Language Proficiency

8 Learner Achievement

In this study, it was important to check that the item difficulty estimates arrived at through the analysis of the survey questionnaires produced, with adjustment for rater severity, sensible estimates for person ability. Firstly, having videos to look at, and having other information about learners (length of study, tests results) enabled a check to be made on the plausibility of the results of the analysis of the items. Secondly, since the range on the logit scale for conference data accounted for by the differences in severity of untrained raters was about equal to the range accounted for by the items (7 logits as compared to 8 logits) it was apparent that following common practice with scales of language proficiency, potential raters would benefit from being able to see calibrated samples on a video showing what level of proficiency was described by the words in the descriptors. In other words, the ability estimates for the video samples offered not only a point of reference during the analysis process, but also the potential to produce a standardisation video for teacher training in relation to the final scale. Finally, the aim of the Swiss Framework project, for which this study was the methodological pilot, was to estimate the range of achievement at cross-over points between educational sectors and provide suitable descriptors of language proficiency as objectives. It was sensible to check that the methodology, or an adaptation of it, was going to provide this.

Decisions on Data Inclusion

In order to arrive at realistic ability estimates for the learners, it was essential to link the two data sets from the questionnaire survey and rating conference. Attempts to do this by analysing the conference data separately and then anchoring together the two complete data sets for the class teacher survey and the conference produced dubious results for two reasons. Firstly a subjective identification of “optimal teacher misfit” was necessary to avoid the length of the logit scale for the video people becoming exaggerated in relation to the values of the items anchored at their questionnaire survey

difficulty estimates. Secondly, the data output from such a fully integrated analysis showed unmistakable signs of the kind of excessive overlap between questionnaires which had been a feature with all analyses which took the data from all seven questionnaires simultaneously.

An alternative approach was therefore taken in which questionnaires were analysed separately. This time, however, the items were anchored at their logit values on the common logit scale. Both learners and teachers were left "floating" in the analysis: that is to say the values for teacher severity from the separate analysis of the conference data were ignored.

One decision to be made was whether to include all 100 teachers who took part in the rating conference, or restrict the analysis to only those teachers who completed the questionnaire concerned. The former approach was adopted for two reasons. Firstly it produced ability estimates for the learners rated on that questionnaire in relation to the perspectives of teachers from all educational sectors. The severity of the teachers who had rated the questionnaire was assessed in relation, not just to other teachers working at the same level, but in terms of a cross section of teachers working at all levels. This increased the coherence of the measurement framework being produced. Secondly, it enabled the full weight of data of the ratings of each video person by 100 teachers at the conference to be taken into account in estimating a learner's ability as well as his/her

teacher's questionnaire ratings. The larger number of counts gave greater precision, lower standard error. This was valuable since (a) these video people were the only common anchoring, so the greater the precision the better, and (b) the fact that the mini-questionnaires had only included between 5 and 7 items to begin with, and the fact that some of these items had by this stage been excluded for misfit or inconsistent interpretation at different levels meant that the number of judgements per rater on each video person was not great.

Because of this desire to include as much of the overall measurement framework as possible, and because of the concern over the small number of items being used to rate each video "anchor person," data was included for video people who had not been rated on the questionnaire concerned, but whose conference mini-questionnaires, through vertical anchoring between the mini-questionnaires, included items which *were* to be found on the questionnaire in question.

Finally, Pronunciation was removed from the analysis of ability estimates for the following reasons:

- The three calibrated Pronunciation items had been added back into the analysis after the other items had been calibrated and had therefore not had an opportunity to influence the difficulty estimates for the other items (though admittedly the effect would have been very marginal).
- Only the top three survey questionnaires or conference mini-questionnaires had pronunciation items left and therefore the ability for the vast majority of learners were going to be estimated without taking account of pronunciation.
- The remaining (high level) pronunciation items showed high misfit anyway in the conference data, probably caused by an interaction between lack of familiarity with Francophone accents and lower quality sound on some of the Francophone video recordings (an unfortunate coincidence).
- The remaining high level pronunciation items now represented approximately 25% of the content area being rated for the video people. However, Pronunciation only represented about 7.5% of the items on Questionnaire E, and between 0% and 3% on the other six survey questionnaires.

The data available for anchoring questionnaires to conference data to establish teacher severity was the judgements of the 100 teachers on the performance of the video persons on the items shown in the Table 8.1. The numbers in the table cells refer to item serial numbers. The prefixes before the learners' names indicate educational sector, as follows:

S Lower Secondary
B Vocational education (Berufsschule etc.)
G Gymnasium (Upper Secondary)
M Adult: Migros Club Schools
V Adult: Volkshochschule (mainly Zürich) and University Language Centre (Lausanne)

Students from the Migros Club Schools (M-) had been rated on all questionnaires whilst adult learners in the state sector (V-) were present only on Questionnaires T2, I and E. These learners together made up the Adult sector. The distribution of questionnaires among the school sectors was more dependent on level. With the exception of two very strong classes which had been rated on Questionnaire T1, all lower secondary classes had received either Questionnaire B (for one year of English) or questionnaires W1 or W2 (for 2 years of English). As regards vocational education, 16–18 year old apprentices had been rated on Questionnaires T1 and T2, with a

couple of classes using Questionnaire I. Two very high level classes of young professionals at a Business school had received Questionnaire E. Finally, the Gymnasium classes (G-), whose approximate level of proficiency was likely to vary most, had been given Questionnaires T2, I and E, depending on how many years English they had had, and what information was available about the standards in the schools concerned. In one or two exceptional cases in Francophone Switzerland in which Gymnasium students had learnt no English in lower secondary, and were now at the end of their first year, Questionnaire W2 was used.

Table 8.1: Common Persons and Common Items Linking Questionnaires to Videos

VIDEO PERSON	B	W1	W2	T1	T2	I
S-Nicole (B)	9 19 20 21 28					
S-Lorenza (B)	9 19 20 21 28					
M-Micheline (B)	33	66				
M-Arlette (B)	33	66				
M-Gertrude (B)	20 28					
M-Marcel (B)	20 28					
M-Renate (W2)		79 80 123	79 80 123	123 159		
M-Rosemarie (W2)		79 80 123	79 80 123	123 159		
B-Marlene (T1)		80	80	132 143 145 158 159	158 159	
B-Pascal (T1)		80	80	132 143 145 158 159	158 159	
V-Florence (T2)		80	80		163 173 200	200
V-Therese (T2)		80	80		163 200	200 209
G-Marina (T2)					173 200	200

Table 8.1 (cont.): Common Persons and Common Items Linking Questionnaires to Videos

VIDEO PERSON	B	W 1	W 2	T 1	T2	I	E
G-Virginie (I)						211 222 236 236	236
G-Christian (I)						211 222 226	236
G-Sibylle (E)							245 259 265 269 273
G-Nils (E)							245 259 265 269 273
V-Yvonne (E)							253 259 266 268 273
M-Doris (E)							253 261 266 273
M-Annemarie (E)							253 261 266 273
M-Eva (E)							253 261 266 273

Taking Questionnaire I as an example, one single video existed for a pair of learners from a Gymnasium in Geneva (G-Virginie and G-Christian). The mini-questionnaire for this video had originally consisted of only 5 items; one of these items (No 238, the fifth on the mini-questionnaire) was a pronunciation item which had been excluded from this particular analysis and No 198 (No 2 on the mini-questionnaire) was a strategy descriptor which had previously been excluded since it had proved to be an unstable anchor. These exclusions reduce to 3 the number of items anchoring the rating at the video conference to the rating in the questionnaire survey in order to estimate learner ability. The three items, Nos 211, 222 and 226 can be found at the point where the column for Questionnaire I intersects with the rows for G-Virginie and G-Christian. Just above the intersection cell for G-Virginie there are some other numbers in the column for Questionnaire I. These are entries for items No 200 and 209 from Questionnaire I. These items had also been included in the mini-questionnaires for the two videos associated with Questionnaire T2 (V-Florence and V-Thérèse; G-Marina). The inclusion of the data from all three of these videos in the analysis of Questionnaire I meant that the anchoring between the rating conference and survey data for Questionnaire I was re-established at 5 common items, and that 5 rather than 2 common persons were available.

The anchoring produced in this manner was in the range which Woods and Baker (1984: 129) cite as adequate (3–10 items). The lack of any video aimed at Questionnaire W1 was, however, a particular problem. FACETS could not accept linking without a common person and so reported two separate subsets of data: that from the questionnaires and that from the rating of video persons on items 79, 80 and 123. The problem was solved by obtaining estimates for teacher severity by averaging the severity for the teachers concerned from (a) the subset values from the W1 analysis, which were based on the conference ratings of the video people concerned on those three items and (b) the severity values for the same teachers from the analysis of Questionnaire W2. These estimates were then used to anchor teacher severity for W1, the conference data then being excluded. This produced a result which made sense and was the best that could be done in the circumstances.

Analysis

Apart from the hiccup with the lack of a video for Questionnaire W1, the analyses were relatively straightforward. What was interesting was a ten-

dency for teachers who had completed the survey questionnaire concerned to "fit" in the analysis, and for "misfit" to concentrate amongst *some* of those teachers who had not completed that particular survey questionnaire. This was only a tendency, but it was very noticeable none the less. It meant that the ability estimates for learners at a particular level were weighted towards the judgements of teachers who had experience of that level and had considered it in detail by rating learners on the 50 items in the questionnaire. But at the same time, these judgements were tempered and put into perspective by the judgements of those other teachers familiar with other levels, or the whole range of levels on the proficiency spectrum, who were capable of making consistent judgements at that level. Those other teachers who were *not* able to offer judgements consistent with the way the majority rated at that level, whether from lack of experience with the level, problems with the video or from prejudice, were excluded. This is quite neat, and in fact reflects the Eurocentres oral assessment model (North 1986; 1991; 1993b) in which judgements made by the class teacher (who knows the students well) are tempered by those of a second assessor (who knows all the levels in the proficiency framework well). This also reflects Cason and Cason's (1984) distinction between high and low sensitive rating around a rater reference point (RRP). Some teachers fitted at all levels, others, even some considered to be assessment experts, misfitted at all

levels other than those to which they were accustomed.

Analysis Runs for Different Questionnaires

In the analysis, it was noticeable that some questionnaires converged very quickly in a couple of analysis runs, with very little misfit, while others required tinkering by excluding teachers and/or learners and up to 8 runs to achieve a result which could be judged acceptable. This was not at all related to the amount of anchoring. The "best" (fewest runs) and the "worst" (most runs) being Questionnaire B and Questionnaire E respectively, the two questionnaires with the most solid anchoring.

In relation to Questionnaire E, the rather noxious issue of judging "optimal teacher misfit" re-emerged, which suggested that the problems of merging two different data sets had still not been fully solved. All analyses in fact showed large residuals on the facet Occasion (School: Conference) during the process of iteration, which indicated that the teachers were in fact reacting differently in the two situations. The position was worst with Questionnaire E. The fact that 32 teachers who did not complete Questionnaire E had to be excluded for misfit, a far greater number than for the other questionnaires, and the fact that many of those excluded were teachers with low level classes suggests that the problem may have been at least partly due to unfamiliarity with such a high level of competence. On the other hand, advanced teachers (native and non-native speaking) who had completed Questionnaire E had in earlier analyses been disproportionately problematic both in terms of misfit and in terms of excessive discrimination (norm-referencing) between learners.

Left to itself without a default value on the number of iterations, Questionnaire E clocked up 570. This is an enormous number and indicates that something funny was going on. In analysing the questionnaires, after pruning out excessive misfit, between 30 and 45 iterations were normally necessary to meet the conventional convergence criteria set: largest outstanding residual less than 0.5; largest value change less than 0.01. The problem may have been caused by Questionnaire E teachers having a greater tendency to change the way they behaved between survey and conference—to change their severity. The effect of removing the 32 misfitting Non-E teachers and of this massive number of iterations was to lengthen the logit scale by about 2 logits. Some of this was probably "good" (removing the Non-E teachers) but some of it was probably bad (change of behaviour by E-teachers). A rerun with a default value of 50 iterations shortened the scale a logit, but

showed distinct signs that all Video People were overestimated in relation to their calibration with the conference data alone. The conventional default of 100 iterations was therefore set for both for this analysis and for the analysis of all the other questionnaires.

The next step was to fine tune for misfit. On Questionnaire E, for example, this meant removing 3 E-teachers and 6 or 7 more Non-E teachers in the process. Then an attempt was made to re-introduce the high scoring learners from the conference data, Yvonne and Beate. These were the learners who had scores over 75%. However, the reinstatement of Beate and Yvonne forced a clear underestimation of the other Video people. The effect of keeping Beate (84% score) would have been to exaggerate the overlap between Questionnaire I with Questionnaire E—making weaker advanced learners come out lower than upper intermediate learners. This demonstrates that high/low scorers *can* distort an analysis and are therefore better excluded whilst difficulty/abilities values are being estimated for items and for other persons. After current exclusions, Yvonne was no longer above the 75% exclusion point but right on it. Further runs with Yvonne, produced a longer scale but teachers, learners and items all now fitted to well under the 1.7. criterion for misfit which had been adopted, so the result including Yvonne was accepted.

The other questionnaires went through a much shorter version of the same process. In the case of Questionnaire B, this involved just two runs, with 4 Non-B teachers, 2 B-teachers, 2 items which were noisy (1.7) and 5 learners being dropped after the first run.

An advantage of having anchored the items to their values on the common logit scale centred on zero, rather than to the values of the items on the individual scale for the particular questionnaire used to rate them, was that the ability values for all the learners were already on the common scale. All that had to be done, therefore, was to import the score files into a word processor, set up a table, sort the table on the logit values and indicate the cut-off points between bands on the scale.

The Learner Scale

The first point that was striking about the scale for learners was that, as throughout the analysis, the scale for learners was much longer than that for items. This is not really that surprising.

Firstly, no questionnaire for the level labelled *Mastery* in the Council of Europe scheme had been included in the survey. This level is intended to represent approximately the level of the Cambridge Proficiency examination (ALTE Level 5). Since there were in fact several classes in the survey who were just taking the proficiency exam at the time, there were good reasons to expect that the stronger learners in those classes (who might get grade "A" at Proficiency) would be calibrated some distance above the highest items, which had been calibrated at the level *Mastery* (intended to be a Proficiency grade "C"). Conversely, no really low level items (e.g. *can say his/her name and address, can say where he/she comes from*) had been included in the survey, and some of the learners being surveyed had had only 60–80 hours English. Since the whole range of proficiency in each class was being surveyed, one could again expect weaker learners in the lowest classes to be calibrated below the level of the lowest items.

Secondly, the items were well targeted because of the pre-testing; none of them even approached scores of under 25% or over 75%. The learners on the other hand could not be pretested. Despite the information about years and hours of study and apparent level, some classes received questionnaires which were pitched a bit high or a bit low for them. This meant that for each questionnaire, including the ones at the top and at the bottom

a full range of achievement from 25% to 75% was always represented. This logically made the person scale longer than the item scale.

The Identification of Mastery Level

The first effect of the calibration of the learners was the discovery of the mistaken identification of the logit range 2.80 to 3.92 as *Mastery*. This was a salutary experience reflecting the fact that the difficulty of descriptors needs to be seen in relation to the way in which they are actually used to rate learners, and not in the abstract. It is perhaps worth recapping the way in which the cut-offs between the levels had been determined. The cut-offs between levels were set at approximately equal intervals, the precise intervals being 0.97 logits right in the middle and 1.10 logits at the two ends. The cut offs were established by fixing equal intervals, looking for natural gaps on the scale, and then by poring over the wording of the descriptors particularly near the cut-offs and comparing this calibration to the intentions of the authors of the source scales. The coherence of the contents of the different levels discussed earlier suggests that cut offs had been fixed successfully.

A detailed look at the items calibrated at the band between 2.80 and 3.92 in relation to the performance of Sibylle, the video person calibrated to this level, certainly did suggest that the relationship between what was per-

formed and what was described made sense. There was no doubt about the fact that she could *express herself fluently and spontaneously, almost effortlessly*, with *little obvious searching for expressions or avoidance strategies*. In terms of other descriptors, not reflected in the holistic scale she could and did *give a clear detailed description of a complex subject* and in a fairly light-hearted encounter with her partner she did *use language flexibly and effectively for social purposes, including emotional, allusive and joking usage.* One had some doubts as to whether she really had *a good command of a broad lexical repertoire* or whether it was true that *only a conceptually difficult subject can hinder a natural, smooth flow of language*, but she had only just scraped into this proficiency band at 3.03 logits and both these two descriptors were above her level at 3.3 and 3.51 respectively with a margin of error of 0.25 logits.

Yet her language did not seem to have the kind of precision which one associates with the label *Mastery* and examinations like the Cambridge Certificate of Proficiency in English. On reflection the albeit limited number of descriptors calibrated at the *next* band seemed to be more appropriate to such a label, and the range of ability represented by Cambridge Proficiency:

> *Has a good command of idiomatic expressions and colloquialisms with awareness of connotative level of meaning.*
> *Can convey finer shades of meaning precisely by using, with reasonable accuracy, a wide range of modification devices.*
> *Can backtrack and restructure around a difficulty so smoothly the interlocutor is hardly aware of it.*

Sibylle's partner Nils (5.05) and Anne Marie (4.74) from a different video did, however, seem to display this quality of precision and idiomaticity. So what had initially been labelled *Mastery* was therefore re-labelled as *Effective Operational Proficiency* and what had initially been labelled *Comprehensive Mastery* was re-labelled *Mastery*.

The Extreme Ability Ranges

That left the question of what to do about really high level language users like the top pair of video people Yvonne (5.89) and Beate (6.36). (The ability estimate for Beate had been obtained from a supplementary analysis.) The next possible band (for which no descriptors were available) was therefore now labelled *Comprehensive Mastery* with a possible category above that,

where Beate was estimated to be, which was labelled "*Ambilingual*" following Trim (1978).

These top two bands for which there are no descriptors, but at which learners were calibrated, are mirrored by two bottom bands for which there was only one descriptor at the level labelled *Smattering*.

Can use some basic greetings; can say yes, no, excuse me, please, thank you, sorry.

Three learners, all from the same class, were rated below the *Smattering* cut off. They included a pair of video people: Micheline and Arlette from the adult education sector. In view of the tendency of Rasch logit scales to distort towards the ends it seemed at least possible that these extreme ranges are a reflection of continuing distortion in the scale, rather than reflecting real differences in ability. Certainly the estimates for Micheline and Arlette appear mean. They are not accepted by the Migros Club schools, who have selected that video as a good example of what they consider to be an Introductory Level (after circa 40 hours). There seem to be three possible explanations for this discrepancy: (a) scale distortion, (b) strict interpretation of *language* proficiency with a down-playing of the communicative success of the simple exchange involved, or (c) a switch in the class teacher severity—a tendency to be stricter on the questionnaire than at the confer-

ence. Since their teacher comes out with almost perfect fit (1.1.) and is in the most lenient 10% of all the raters at a logit value of -1.64 (minus = lenient), but is considerably less lenient on the conference data as a whole (-0.64), explanation (c) seems unlikely.

This leaves explanations (a) scale distortion and explanation, and (b) that there is a difference of perspective. Explanation (b) is plausible: the result could well be a correct reflection of what was rated. The majority of the teachers dominating the construct were from the lower secondary school sector. The result could reflect schoolteachers being tough on learners struggling gamely with three or four word sentences and non-linguistic resources.

Looking at the same issue at the top of the scale, there are two teachers involved. Beate's teacher at the University of Lausanne Language Centre has good fit (1.2). With a severity of 1.43 on Questionnaire E and a severity of 0.64 for the whole conference data, she is tougher on advanced students than she is overall. If anything, therefore, she would seem to tend to err towards being too strict on her own students. Therefore these high levels do not appear to be exaggerated by switch in severity (Explanation c). That would seem to confirm the existence of a very high level of *Comprehensive Mastery*, in effect *Ambilingualism*, amongst some recipients of Language Centre courses at Lausanne University. It is worth mentioning that these two learners were not students at the university, but rather administrative employees fluent in several languages.

The second teacher, from a Business school in Ticino, was a very strict rater. In fact she is *the* strictest rater with a severity value of about +3 logits on both Questionnaire E and the whole conference data. She is "rather noisy" at 1.4 with 2 for standardised residuals, but both are within the conventional criterion for unidimensionality. As with all three teachers discussed here, the three learners are bunched closely together; there is no hint of norm-referencing. The result seems possible.

So one possible and one plausible result. Could it be that the Linacre model overcompensates slightly for teacher severity/lenience, and that this effect is magnified at the two ends of the scale, as Rasch results always seem to be magnified at the two ends of the scale? Could it also be that such an effect is magnified for the negative logit values (c.f. Micheline and Arlette) as suggested by Lord (1983) and Warm (1989)? An investigation of the effects of Rasch logit scale distortion on logit values in a three facet (Item,

Person, Judge) analysis of a full data set would be an interesting area for further research.

Achievement in Educational Sectors

It was not the aim of this study to identify the range of achievement of Swiss learners, but having now constructed and interpreted a scale of descriptors, and found a way to take account of teacher severity in estimating ability values for learners, it was possible to give an overview. Naturally the validity of any statements would be limited by a number of issues. Firstly, the descriptors reflect teacher assessment. Other perspectives from test results or discourse analysis of (classroom) speech would be necessary for any detailed picture. Descriptions of proficiency drawn from the descriptors probably reflect the way teachers think about their learners' proficiency as much as that proficiency itself. Secondly, only those aspects or levels of proficiency for which descriptors were calibrated can be described; gaps in coverage can be due to a variety of reasons. Thirdly, one should be cautious about conclusions drawn on the basis of a single cross-sectional survey on the basis of a Rasch calibration, especially when corrective measures had to be taken for teacher norm-referencing and distortions caused by the measurement model. Finally, whilst attempts were made to recruit teachers from particular regions and sectors to get balanced coverage, there are very defi-

nite limits to the extent to which one could say that the results are representative.

Having said that, the results certainty paint an interesting and really quite plausible picture, as can be seen from Figures 8.1, 8.2 and 8.3, which report results in terms of achievement for years of study in the different educational sectors. The letters heading the columns on the charts refer to educational sectors and the numbers refer to the number of years of English learners were *thought* to have had by their teachers. Thus:

S1	Lower Secondary, up to 1 year of English
S2	Lower Secondary, up to 2 years of English
S3	Lower Secondary, up to 3 years of English
G1-6	Gymnasium 1,2,3,4,5,6 years of English
B1-5	Berufsschule /Vocational Training 1,2,3,4 or 5 years
M1 etc.	Migros Club Schools 1, etc. years of English
M I	Migros Intermediate (years of English not known)
M U	Migros Upper Intermediate (years of English not known)
V I	VHS (Volkshochschule) / University Intermediate
V U	VHS (Volkshochschule) / University: Upper Intermediate
V A	VHS (Volkshochschule) / University: Advanced
V Pro	VHS (Volkshochschule) / University: Proficiency

In Figure 8.1, approximately 470 learners are represented, each number representing a learner. This is only just over 50% of the original 945, but one has to bear in mind that approximately 40% were immediately lost by the necessity to remove the 1st and 5th learners from nearly all classes in order to counter the effects of excessive norm-referencing. In fact 140 of the original 151 classes are still represented and the learners can be taken to be representative of the classes concerned.

Two things are immediately striking about the chart. Firstly, there is a wide range of level represented by the posited achievement for each year of this sector, a range which stays relatively stable. Secondly, there is quite remarkable coherence in the progression within each sector. The break in the Migros Club school progression at years 5–6 reflects a fundamental switch in the Migros learners from those who have been following a syllabus of Migros proprietary materials, and a mixture of (a) exam classes and (b) conversation groups. The latter are generally *femmes d'un certain âge* who may have been meeting together once a week for 10 years or more. The extraordinarily wide range of achievement of people who are all apparently in

classes which their teachers chose to label upper intermediate is therefore due to the fact that both Cambridge First Certificate classes and such long-standing conversation groups were both labelled upper intermediate.

Lower secondary learners appear to cover the same range of level whether they have had one year or two years of English, but the mean weighting has clearly switched after two years when the majority of learners are reaching *Waystage* or beyond. Upper secondary learners (G = Gymnasium) appear to make slightly faster progress, as one might expect. Those with no previous English seem to reach *Breakthrough* and *Waystage* after one year, and half have reached *Threshold* or above after two years. The position for the vocational schools is more confused. Here one is dealing with learners who have an hour or so a week in very mixed ability classes, since apprentices whose employers allow them to come on Tuesdays are the Tuesday class. Those learners with three years of English show an achievement range comparable to that of lower secondary children with three years. The abrupt change in the nature of the progression at B5/UI is really a classification error. The learners in B1-B4 are apprentices, whereas those in BUI and BAd are the young professionals studying at a Business school and taking Cambridge Proficiency, who were referred to earlier.

The learners marked with an X are from two classes belonging to the same teacher, who is clearly behaving in a different way to everybody else.

That teacher phoned up to ask whether she should be rating her learners in relation to each other, their relative position in the class, or in relation to their position in the overall imaginable range of proficiency. She had great difficulty coming to grips with the idea of rating in relation to a defined criterion (the descriptor) and using a 0–4 scale to do so. She persevered and remarked at the conference that she felt she might have rated her questionnaires incorrectly. She also misfitted considerably on all video ratings except those at advanced level, where she taught. There were 6 other learner calibrations which strained belief, coming out well above or below the main learner scale for a questionnaire. All 6 *might* reflect reality, but if they do, those learners are not typical of their sectors, and thus do not add to a picture of typical achievement which might help in the selection of appropriate standards.

The most regular pattern is to be found among the small numbers of learners in the Volkshochschule / University language centres. Here native speaker professional EFL teachers used level labels rather than years of English to give information about their learners before the survey. The results confirm that people were by-and-large the level their teachers said they were.

To check the regularity of the patterns of progress, means, medians and standard deviations were calculated. A summary of the result is shown in Figure 8.2. This time the shading represents the area covered by one standard deviation from the mean, and the black square represents the median: the mid point in terms of the range of proficiency covered. The shaded pattern is similar to the previous chart, but there are various anomalies caused by sampling. For example, it is scarcely credible that lower secondary learners would reach *Vantage* after three years. This effect is caused by the small size of the sample for S3 and the fact that the SDs are calculated from the mean, which is near the top of the range of level covered by the sample. The progression in the Gymnasia also shows a hiccup: a relatively large sample for G2 with a large proportion reaching *Threshold* and *Threshold Plus* is followed by a small sample for G3 with the majority only at *Waystage Plus*. These distortions clearly limit the validity of this chart. However, since the first chart was limited to the learners who happened to be in the survey, the second chart did at least provide pointers to suggest ranges where no learners had actually been calibrated, but where one might expect learners to be found.

Figure 8.3 represents a compromise between the two already discussed. Whereas the first two charts are objectively based on data, the first on logit calibrations, the second on medians and standard deviations from the mean, this third chart is conjectural and seeks to show the *likely range of achievement* in the different sectors by smoothing the patterns presented in the previous two charts.

The Development of a Common Framework Scale of Language Proficiency

Learner Calibrations from Two Data Sets

One of the problems discussed in Chapter 5 was a possible tendency for the teachers to use descriptors in a different way depending on whether they were rating learners in their class or rating videos of other learners at the conference. A totally independent comparison of the way descriptors were used on the two occasions is difficult, as it is only through the link provided by the conference that adjustments for rater severity can be made in order to calculate ability estimates for the learners. A second problem concerned linking the two data sets from the conference and from the survey. Therefore it is interesting to compare the calibrations for the video people arrived at through the analysis of assessment data from the two different contexts, as is done in Table 8.2.

Table 8.2: Learner Ability Values from Two Data Sets

Level	Video Person	Conference	Survey
M+	V-YVONNE		5.89
M	G5-NILS		5.06
	Mad-ANNEMARIE		4.74
EOP	G5-SIBYLLE	3.05	3.03
	Mad-EVA		2.94

V+		Mad-DORIS	2.19	1.94
V	*	G4-VIRGINIE	.55	1.39
	*	G4-CHRISTIAN	.28	.98
T+	*	VIn-THERESE	-.80	-.01
	*	G4-MARINA	-.36	-.22
	*	M3-RENATE	-.17	-.57
T		M3-ROSEMARIE	-.57	-.78
		B3-MARLENE	-.92	-.94
W+	*	VIn-FLORENCE	-2.49	-1.59
		S1-NICOLE	-3.70	-3.32
B		B3-PASCAL	-3.40	-3.34
	*	M3-MARCEL	-2.01	-3.48
Tour		S1-LORENZA	-4.71	-4.36

The two assessment contexts contrasted in Table 8.2 are: *Summative Assessment* (Conference) of one 6–12 minute videoed performance by 100 teachers in relation to the performances of the other video people; *Teacher Assessment* (Survey) on a wide range of tasks (represented by the descriptors) by just the class teacher in relation to other learners in the same class. The two sets of calibrations compare as shown in Table 8.2.

The Spearman rank correlation is 0.94 (n=14; p = < 0.01). The Pearson correlation of the actual scores (suspect with a sample size under 30) is 0.96 (n=14; p = <0.001). However, these quite impressive correlations do disguise some noticeable differences. 7 of the 14 learners whose calibrations can be compared would be placed at a different band on the scale. These 7 are marked with a asterisk above. In most cases the difference is easily explained by the circumstances of the video recording.

VIRGINIE & CHRISTIAN were two French-speaking learners from Geneva who gave a relatively poor performance explaining aspects of the plot of "The Importance of Being Ernest" in a very stiff setting, on a video cassette with really terrible audio quality. Even their own teacher remarked at the conference how poor the performance was in relation to what they could do normally. A very plausible explanation is that we have two justifiably different assessments: the one of what these two learners could be expected to do on the basis of their coursework, the other, how they were rated on a poor performance which was difficult to hear.

THERESE & FLORENCE were two adult French-speaking learners from Lausanne, again on a video tape which was difficult to listen to. It is

noticeable that they are both downgraded heavily when the conference data is considered alone.

MARINA shows very little change, she just happens to move over the cut-off point because she is right on it. This shift is within the standard error.

RENATE is a similar case to MARINA. She was assessed as Eurocentres Level 5.5 on Itembanker just into *Threshold Plus*. At the conference, benefiting from the "bonus" for an excellent performance, she makes it over the cut-off, again into *Threshold Plus.* Her teacher was the most lenient at the conference, so FACETS has adjusted her rating in the survey downwards. The possibility that FACETS might overestimate this adjustment was discussed in earlier in the chapter, but the Itembanker data suggests that she really is exactly on this cut-off, so there is no reason to think that the adjustment is radically wrong.

MARCEL is the only case where there is really cause for concern. His teacher is the same teacher as for GERTRUDE. This teacher showed a tendency to misfit both on certain questionnaires and in separate analysis of the conference data. This misfit (inconsistent rating) is demonstrated by the fact that he rated GERTRUDE in the survey *higher* than MARCEL although it is blindingly obvious from the video that Marcel is much, much better than she is. On the other hand he benefits from the good performance bo-

nus in the video, especially as he helps GERTRUDE express herself all the time. It is therefore very likely that MARCEL has for some reason been quite simply under-rated by his teacher.

In conclusion, therefore, the comparison shows that both kinds of assessment (teacher assessment / summative assessment) have their drawbacks. This in turn suggests that a fully effective assessment approach would combine both, as is the case in the English National Curriculum and in Eurocentres. Differences in the ability estimates for the video people under the two assessment conditions are attributable either to the particular performances rated on the videos, or to particular teachers, and show no grounds for supposing that the descriptors are interpreted in such a way that they cannot be used for both types of assessment.

Concurrent Validity

Unfortunately, very little test data was available to give an independent assessment of the ability of the learners rated in the survey. By coincidence, 53 of the adult learners at Migros club schools who were included in the survey were also among 109 classes who took Eurocentres Itembanker tests in a separate Migros study. Of these 53, many were excluded because of teacher norm-referencing or extreme scores, but ability estimates were obtained for 37. Unfortunately, 9 of those 37 were in classes of a teacher who gave the Itembanker tests 4 months after the survey, so the available reliable data was limited to 25 learners from 9 teachers, including two of the video persons Renate and Rosemarie. 18 of these 25 learners were placed between Levels 4 & 6 on the Eurocentres Scale (thought to be *Threshold* and *Threshold Plus*).

The correlation between the survey result and the Itembanker result is not particularly high at 0.317 (Spearman ranking: not statistically significant) or 0.448 (Pearson: $p > 0.05$) on the actual logit estimates of the two assessments (survey; Itembanker). However, one would not expect a very high correlation for two reasons. Firstly, a low correlation between a teacher assessment of communicative performance and a test of linguistic knowledge is not surprising in an acquisition-poor environment. Secondly, since 72% of the sample are on the three middle bands (4–6) of the Eurocentres scale, we have a truncated sample, which again leads to a lower correlation (Nunally 1978). When results on the survey were plotted linearly against the Eurocentres scale bands reported by the Itembanker program, the results were, despite the low correlation, encouraging.

The cut-off for Eurocentres Level 4 (thought to be *Threshold*) came out at almost exactly -1.23 on the logit scale for the survey, the cut off established for *Threshold*. Eurocentres Level 8 came out at 1.67, which is more or less the cut-off for *Vantage Plus* (1.74) with which it had been identified. Eurocentres Level 6 was 0.375. By a process of simple arithmetic, Eurocentres Level 7 (if considered equidistant between 6 and 8) would then come out as about 1.02. If Eurocentres Level 7 is 1.02 and Level 6 is 0.375, then Eurocentres Level 6.5 (equated to Cambridge First Certificate "C" Pass: North 1991; 1994 and thought to be *Vantage)* would come out at 0.70, almost exactly on the cut-off established for *Vantage* (0.72).

Therefore, on this admittedly very thin basis of results for 25 learners, the relationship between the survey scale cut-offs and the Eurocentres scale bands reported from a Rasch based item bank was virtually spot on, as summarised in Table 8.3.

Table 8.3: Concurrent Validity for 25 Learners

Levels	Associated Eurocentres Level	Survey Level Cut-off	Plotted from Eurocentres Itembanker

V+	Level 8	1.74	1.67
V	Level 6+	0.72	0.70
T+	??	-0.26	-
T	Level 4	-1.23	-1.23

Reading backwards from the Survey scale to the Eurocentres scale, it would appear on the same basis that *Threshold Plus* is not associated so neatly with a Eurocentres Level, coming out between Level 5 and Level 5+.

Based on such a small sample, this result should not be taken too seriously, but it is, to say the least, encouraging. It suggests that at least at the middle range of the logit scale, where it is accepted in the literature to be linear, the logit scale produced in this study and the Itembanker logit scale produced in a previous PhD study (Jones 1993) can be related to one another in a linear fashion. It also supports the decisions on cut-off points on the logit scale made in this study.

The Development of a Common Framework Scale of Language Proficiency

9 Conclusions

This study has developed and tested a methodology to address a practical, felt need. There is a widespread development of common framework scales (e.g. in the UK: English-speaking Union Framework Project, English National Curriculum, UK National Language Standards; in Europe: Council of Europe Framework Project, Association of Language Testers in Europe; LANGCRED). A methodology for the development of such instruments has until now been lacking. This study demonstrates *one* way in which such an undertaking can be done in a principled fashion.

Project Results

It can be claimed that the scale development has taken account of the main problems with common framework development discussed at the end of Chapter 1. Firstly, by working within and contributing to the set of descriptive categories being developed by the Council of Europe Common Framework authoring group of John Trim, Daniel Coste and the author, it has been possible to relate the categories to theoretical models of language use as described in Chapter 2. Secondly, by working interactively with over 50 teachers in the series of workshops described in Chapter 4 and then calibrating the descriptors in relation to the judgements of 100 teachers in relation to their students, it has been possible to keep both the categories employed and the descriptors defining them user-friendly. Thirdly, by investigating the issue of variable interpretation in different educational sectors and language regions (Differential Item Functioning) and in relation to different assessment occasions it has been possible to establish a degree of context-independence. A measure of the consistency of interpretation is represented by the item quality hierarchy presented at the end of Chapter 6. Approximately 80% of the descriptors show no significant variation by context, and a set of “excellent items” with a very high degree of context freedom has been identified. Fourthly, the situation of the descriptors on the scale is objective to the extent that the collectivisation of subjective

judgement can ever be objective. The scale meets Thurstone's and Thorndike's requirements for a valid measurement scale. Finally, both a pragmatic and an empirical solution have been found to the vexed question of the number of levels which should be presented in a common scale. A twin scale approach (North 1992a, 1992b) is proposed: narrower levels at equal intervals of approximately 1 logit grouped into broader Common Reference Levels.

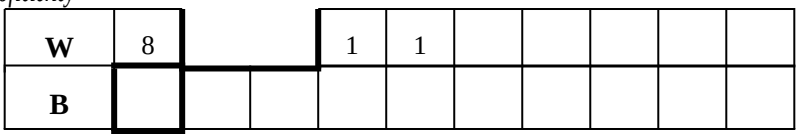
The scale can therefore be said to represent the mapping of a set of criterion statements onto the dimension of progress in spoken English on the basis of the judgements of reasonably representative potential future users. Formulations are not interdependent. Each descriptor is independently calibrated, yet the wordings none the less show remarkable coherence. Several common criticisms of scales of language proficiency discussed in Chapter 1 are answered in this way. Though previous work in the field is built on systematically, conventions are not copied unthinkingly (North 1992a); the provenance of descriptors is stated (Brindley 1986); the scale *can* be broken up into criterion statements to which one can say Yes or No (Skehan 1984); progression is achieved without juggling qualifiers like *a few, some, many (*Alderson 1991a); the argument is *not* circular (Lantolf and Frawley 1988, 1992); the descriptors are not all subsumed into holistic paragraphs in which the co-occurrence of features are counter-intuitive (Skehan

1984: 217; Brindley 1986: 56; Van Ek 1987: 24; Fulcher 1987, 1988); the scale can genuinely be claimed to represent somewhat more than just what proficiency *might* look like (Clark 1985: 348). One can in fact claim a degree of the *a priori* validity (Fulcher 1993/6) appropriate to a common framework. Other data-based forms of development may be more appropriate for developing defined rating scales to be used with particular tests (e.g. see Fulcher 1993; Upshur and Turner 1995; Brindley 1998) and other ways will no doubt be discovered to develop common framework scales, but this book has described the development of *one* methodology to do it.

It is interesting to compare the extent to which the position of descriptors on the scale produced in this study relates to their position on the scale from which they originate. As mentioned in Chapter 7, this was done informally as part of the process of establishing the cut-offs between levels on the scale. Approximately one third of the 212 calibrated descriptors originate solely from Eurocentres, and approximately two thirds originate at least partly from a Eurocentres formulation. Therefore it is interesting to plot the position of descriptors shared by the two scales, as in the Figure 9.1. The relationship is a correlation of 0.884.

Figure 9.1: Interaction Descriptors from the Eurocentres Scale

Eurocentres Levels										
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
M										
EOP										
V+										
V										
T+							1		1	
T		1	2			1				
W+	1			1	1					



The shaded, boxed area represents the relationship thought to exist between the two scales. As can be seen there is a very definite pattern. That 8 Eurocentres Level 1 descriptors are calibrated at *Waystage* rather than *Breakthrough* is almost certainly caused by the fact that the kinds of simple real-life tasks referred to will tend to be considered easier in the acquisition-rich environment of a stay abroad at a Eurocentres. In addition, the survey values at *Breakthrough* are mainly based on ratings by lower secondary school teachers, and, as pointed out whilst discussing DIF in Chapter 6, there was a tendency for these items to be rated as harder for teenagers than for adults. The 4 items at Eurocentres Levels 4–5 (*Threshold*) placed at *Waystage* and *Waystage Plus* reflect the phenomenon discussed in Chapter 6 in relation to Description items from the ASLPR: the teachers interpreted some description tasks to be simpler than sometimes intended by scale writers. The placement which is really odd is the Level 9 item calibrated at *Threshold Plus*. This was the following descriptor edited from Eurocentres Level 9 and from Level 3 scales from the FSI family: *Pronunciation is clear and intelligible even if a foreign accent is sometimes evident and occasional mispronunciations occur.* This

just goes to show that something rather odd was happening with Pronunciation, as discussed in Chapter 6.

Replication in Year 2

The second year of the Swiss research project was organised as a replication study, extending the survey to Listening and Reading, and to French and German as well as English.

Project Design

The project followed exactly the same three phases that had been established in Year 1:

Creating a Descriptor Pool for Listening and Reading. A similar classification scheme for communicative activities, strategies, qualitative aspects of proficiency and socio-cultural competence was again employed. Once again editing produced a pool of approximately 1,000 descriptors.

Qualitative Validation: Consultation with teachers through workshops. 14 workshops with about 150 teachers were conducted, with sorting tasks as in Year 1. During these workshops, the rejection rate was considerably higher than in Year 1. Both very global statements and statements trying to define linguistic qualities of texts which could be understood tended to be unpopular with teachers. More concrete information about activities tended to be preferred.

Quantitative Validation: Scaling descriptors through teacher assessments. The same range of level as 1994 was covered with four questionnaires, with a fifth very high level questionnaire which in the event did not yield enough data for a satisfactory analysis. 61 of the 170 items employed on the four questionnaires which could be analysed provided anchoring back to the 1994 English survey. Parallel scale construction analyses were run, one anchoring the 61 items from Year 1 back to their 1994 values in order to link the two analyses onto the same scale, and the other allowing the 1994 items to float and establish new values. The Wright and Stone (1979) anchor checking technique was again employed to check stability of

anchor values between forms as in 1994. It was also exploited to check the stability of the difficulty estimates of the descriptors common to the 1994 and 1995 surveys.

Following Bejar (1980), sub-analyses were also run for the three main content strands: interaction, listening and reading to see if there were dimensionality problems. Reading appeared not to fit a construct dominated by the overlapping concepts of speaking and interaction, despite the fact that Rasch fit statistics did not indicate this. This conclusion was reached since reading items showed:

a. A consistent tendency to Differential Item Functioning (DIF): to show apparently significant differences in the difficulty values obtained from different groups of learners.
b. A different slope to the scale when a separate analysis was plotted against the main construct.
c. A difference larger than standard error in difficulty estimates when anchored to the 1994 construct or when analysed separately.

Reading was therefore analysed separately. The resulting reading scale was anchored to the main scale by analysing reading together with listening (listening + reading = Reception). The 37 calibrated listening items were then used as anchor items in order to equate the two subsequent scales. The

relationship between the two scales proved in fact to be linear with a correlation of 0.98, but the listening items were calibrated 0.31 logits lower when in company with Reading. The reading items were therefore equated to the speaking-interaction-listening scale by increasing their values on the reading & listening scale by 0.31.

Replication scale values for descriptors

The central aim of the 1995 survey was to see if the 1994 scale values for descriptors would be replicated in a survey focused mainly on French and German. The difficulty values for the items in the 1994 construct (spoken interaction & production) proved to be very stable. Only eight of the 61 1994 descriptors reused in 1995 were interpreted in a significantly different way—i.e. fell outside the Wright and Stone's 95% criterion line. After the removal of those eight descriptors, the values of the 103 listening & speaking items used in 1995 (now including only 53 from 1994) correlated 0.99 (Pearson) when analysed (a) entirely separately from 1994 and (b) with the 53 common items anchored to their 1994 values. This is a very satisfactory consistency between the two years when one considers that:

1. The 1994 difficulty values were based on judgements by 100 English teachers, whilst in 1995 only 46 of the 192 teachers taught English, and only 20 of them had taken part in 1994. The ratings dominating the 1995 construct were therefore those of the French and German teachers.
2. The questionnaire forms used for data collection in 1994 and 1995 were completely different in terms of both content and range of difficulty with 4 forms in 1995 covering the ground covered by 7 forms in 1994.
3. The majority of teachers in 1995 were using the descriptors in French or German. Therefore it is possible that the problems with the eight 1994 descriptors may have been at least partly caused by inadequate translation.

Replication of scale values for video performances

The proficiency values estimated for the learners shown in the video samples showed a high degree of stability when reused in the following year.

The values for seven English videos reused in 1995 correlated well to 1994 values (r = 0.97; n= 7). In addition, a certain number of videos were used to introduce teachers to rating learners directly onto the scale of levels at a conference in September 1996. A technique combining holistic and analytic rating developed by North (1991) was used with a scale and grid constructed from the descriptors calibrated in the project. After short initial training, groups of in each case approximately 25 raters for English, French and German (of whom approximately two thirds had taken part in the survey) rated 3 video performances onto the scale. Of the total of 9 video recordings rated, 7 had been previously calibrated in the survey. Logit scores for these conference ratings calculated by anchoring the rating scale step values to the logit values of the cut-offs given in Table 7.2 correlated well with the values obtained in the survey itself (r = 0.96; n = 7). The relatively high correlation disguises the fact that only 4 of the 7 video extracts were actually rated onto exactly the same level in both contexts and not many performances are involved, but the result is certainly encouraging.

Differential Item Functioning

There was considerably more DIF in the second year, as discussed in the article reporting on the project (North and Schneider 1998) and outlined in detail in the project report (Schneider and North 1999). The results for educational sectors and language regions were similar to those discovered in the English survey. Patterns for target language and mother tongue were difficult to discern. There certainly did not appear to be any curriculum effect in relation to the pedagogic cultures of the three target languages. An analysis of the difficulty values from ratings of people teaching their own mother tongue (approximately 25% of the total) compared to the difficulty values from ratings of people teaching what for them was a foreign language showed full stability for the dominant spoken interaction/production construct. Just a handful of listening and reading items showed statistically significant variation at just over the 5% level. There was a suggestion here that the non-native speaker teachers might have thought of "understand" and "follow" as synonyms, whereas the native-speakers may have meant rather more than just understanding the words and propositions involved, for example with regard to literature. What DIF was occurring in relation to individual descriptors balanced itself out in the aggregate results. The analyses quantifying the effect of the "facets" target language, educational sector, language region and mother tongue showed no statistically significant values.

Conclusions in Relation to the Rasch Model

What has been unique in this current study is the extension to judgemental data of an item-banking methodology, with a over-lapping "missing-data" design produced by a series of tests/questionnaires covering the full range of ability. This adaptation of item-banking methodology ran into some quite serious problems. However, these were overcome and the methods adopted to correct for them may be of value to other researchers in the field. Statements about the Rasch model in the literature remind me a little of ancient soothsaying: nothing is ever actually incorrectly stated, but it turns out that some direct experience of Rasch analysis is a great advantage in interpreting apparently simple statements—and noticing caveats. Woods and Baker, in their introduction to the Rasch model write:

The Development of a Common Framework Scale of Language Proficiency

> "It seems to us that, provided the use of the Rasch model is tempered by common sense and experience—as *any* item analysis ought to be—and provided it is not intended to create a "once-and-for-all" bank of items whose properties are expected to remain constant over time, the Rasch model can be a useful tool for identifying a set of items which cover the part of the ability range in which the tester is interested." (Woods and Baker 1984: 131)

At the time of first reading this rather sophisticated sentence I focused, where I expect other readers will tend to focus, on the statement about the question of stability over time, which I later discovered was a response to concerns raised by Goldstein (1981) and Tall (1981). This led to the suggestion to develop a illustrative descriptor-bank rather than a scale for the Council of Europe Common European Framework, since a Rasch analysis offers the opportunity to identify and adjust for the development of curriculum practice over time and expand the bank.

However, after an analysis which turned out to be considerably more complex than had been envisaged, my eye was drawn by the statement about tempering analysis results with common sense and experience, but above all by the statement about identifying "a set of items which cover the *part* of the ability range in which the tester is interested." This is at the heart

of the problem of excessive overlapping between forms; this is at the heart of the excessive norm-referencing on the part of the major of teachers. Cason and Cason (1984) talk of raters having a "rater reference point" (RRP: a focus at the level they are familiar with) and "sensitivity." A rater with high sensitivity is a rater who discriminates well in the immediate vicinity of his/her RRP; a rater with low sensitivity is one who rates well over the whole continuum, not necessarily any better near his or her RRP. The way the majority of the 100 teachers rated suggests that most teachers have "high sensitivity," or to put this negatively, a narrow horizon. Within the visible environment—the breadth of which is determined by their breadth of vision—they can rate their learners, though they do so to varying degrees of strictness/leniency. However, because what is over the horizon is less immediate, less real, they tend to "overuse" the measurement scale in relation to the segment in focus.

On careful re-reading, Woods and Baker seem to be saying that the Rasch model itself operates in a similar fashion. It will tell you what other items are in the same part of the ability range as those that you have, perhaps calibrating really accurately in the range from -2 to +2 logits as suggested by Camilli (1988: 231). But when one chains together test forms across a number of ability ranges, tackling the whole ability spectrum, one is going to run into problems, as Jones (1993) discovered and as this study discovered.

Yet the problems do appear to be soluble, at the price of excluding a proportion of the data on learners. In this respect it was fortunate that the instructions for the choice of learners had been so directive: one was able to define what one was excluding. The product is a scale of language proficiency which, as discussed in Chapter 7, demonstrates such a degree of content coherence that one can have some considerable confidence that the technical problems encountered were in fact overcome. As a result, the scale probably does represent the nearest thing to a linear scale of proficiency which it is currently possible to produce.

Conclusions in Relation to Descriptors

The essentials of a valid measurement scale were discussed in Chapter 3, and several of the points raised related to the formulation of descriptors. Chapter 4 described the process of forming a descriptor pool and discussed

teachers' reactions to the descriptors in the pool and Chapter 6 discussed in some detail the weaknesses of individual descriptors identified in the course of the analysis. The overall impression gained on what makes a descriptor work is summarised in the list of points below:

Positiveness

It is a common characteristic of assessor-orientated proficiency scales and of examination rating scales for the formulation of entries at lower levels to be negatively worded. It is more difficult to formulate proficiency at low levels in terms of what the learner can do rather than in terms of what they cannot do. But if levels of proficiency are to serve as objectives rather than just for screening candidates, then positive formulation is desirable. It is sometimes possible to formulate the same point either positively or negatively, e.g. as shown in relation to *Range* of language in Table 9.1.

An added complication in avoiding negative formulation is that there are some features of communicative language proficiency which are not additive, that is to say the less there is the better. The most obvious example is what is sometimes called *Independence*, the extent to which the learner is dependent on (a) speech adjustment on the part of the interlocutor, (b) the chance to ask for clarification, and (c) the chance to get help with formulating what he wants to say. Often these points can be dealt with in pro-

visos attached to positively worded descriptors, for example: *can generally understand clear, standard speech on familiar matters directed at him/her, provided he/she can ask for repetition or reformulation from time to time*; or: *can understand what is said clearly, slowly and directly to him/her in simple everyday conversation; can be made to understand, if the speaker can take the trouble.*

Table 9.1: Positive and Negative Descriptor Formulation

Positive	Negative
- has a repertoire of basic language and strategies which enables him or her to deal with predictable everyday situations. (Eurocentres Level 3: certificate) - basic repertoire of language and strategies sufficient for most everyday needs, but generally requiring compromise of the message and searching for words. (Eurocentres Level 3: assessor grid)	- has a narrow language repertoire, demanding constant rephrasing and searching for words. (ESU Level 3) - limited language proficiency causes frequent breakdowns and misunderstandings in non-routine situations. (Finnish Level 2) - communication breaks down as language constraints interfere with message. (ESU Level 3)
- vocabulary centres on areas such as basic objects, places, and most common kinship terms. (ACTFL Novice)	- has only a limited vocabulary. (Dutch Level 1) - limited range of words and expressions hinders communication of thoughts and ideas. (Gothenburg U)
- produces and recognises a set of words and short phrases learnt by heart. (Trim 1978 Level 1)	- can produce only formulaic utterances lists and enumerations. (ACTFL Novice)

Definiteness

Descriptors should describe concrete features of performance, concrete tasks and/or concrete degrees of skill in performing tasks. There are two points here. Firstly, the descriptor should avoid vagueness, like, for ex-

ample: *Can use a range of appropriate strategies*. What is meant by strategy? Appropriate to what? How should we interpret "range?" The problem with vague descriptors is that they can read quite nicely, but an apparent ease of acceptance can mask the fact that everyone is interpreting them differently. Secondly, since the 1940s, it has been a principle that distinctions between steps on a scale should not be dependent on replacing a qualifier like "some" or "a few" with "many" or "most" or by replacing "fairly broad" with "very broad" or "moderate" with "good" at the next level. Distinctions should be real, not word-processed and this may mean gaps where meaningful, concrete distinctions cannot be made.

Clarity

Descriptors should be transparent, not dense, verbose or jargon-ridden. Apart from the barrier to understanding, it is sometimes the case that when jargon is stripped away, an apparently impressive descriptor can turn out to be saying very little. Secondly, they should be written in simple syntax with an explicit logical structure. Double-barrelled descriptors (can do X but cannot do Y) as found in ELTDU (1976) and ALTE (1994) appear to be less easy for teachers to relate to. This may be because they combine a positive with a negative and only one may be true of the person concerned. It

may be because of the breach of the *Positiveness* requirement. Fundamentally, however, any two clause sentences linked by "and" or "but" should be looked at carefully to see if they should be split; three clause sentences appear to be definitely too complex.

Brevity

There are two schools of thought. The one associated with holistic scales, particularly those used in America and Australia, tries to produce a lengthy paragraph which comprehensibly covers what are felt to be the major features. Such scales achieve "definiteness" by a very comprehensive listing which is intended to transmit a detailed portrait of what raters can recognise as a typical learner at the level concerned, and are as a result very rich sources of description. There are three disadvantages to such an approach: Firstly, no individual is actually "typical." Detailed features co-occur in different ways. Secondly, a descriptor which is longer than a two clause sentence cannot realistically be referred to during the assessment process. Finally, teachers consistently seem to prefer short descriptors. In the workshops, teachers tended to reject or split descriptors longer than about 20 words, approximately two lines of normal type. Interestingly, Oppenheim (1966/1992: 128) also recommended up to approximately 20 words for opinion polling and market research.

Independence

Two further advantages of short descriptors are that (a) they are more likely to describe a criterion behaviour, that one can say Yes or No to the question whether the person can do this, and (b) that consequently they can be used as independent criterion statements in checklists or questionnaires for teacher continuous assessment and/or self-assessment. This kind of independent integrity is a signal that the descriptor is actually saying something rather than having meaning only relative to the formulation of other descriptors on the scale, and therefore broadens the range of assessment formats in which the descriptor could be used.

Areas for Follow-up and Further Research

There are a number of issues which could benefit from follow up and from future research. First of all it would be an exaggeration to say that the scale produced has been fully validated. There is a considerable amount of *a priori* validation which answers many of the criticisms of scales of language proficiency in the literature. There is also *concurrent* validity from three follow up projects using the descriptors for self-assessment, which have each demonstrated a remarkable stability in the scale values of descriptors, as reported briefly below. But there is as yet no large scale *a postieri* validation to prove that the scale "works" as an assessment scale when the descriptors are used operationally.

It is clear that one follow up needed is controlled experimentation with the creation of continuous assessment checklists and summative assessment criteria grids for local systems, with trialling and analysis. What might be particularly interesting would be to use the communicative activity descriptors as the specification for the construction of oral assessment tasks, and then to use the descriptors on strategies and qualitative aspects of language use to rate performance in them. Three projects taking place in Switzerland are working in these areas:

a. The further elaboration of and experimentation with the Language Portfolio by a network of Swiss teachers as part of the Council of

Europe project aiming to launch the European Language Portfolio in 2001. Reports are given in Schärer (1999).

b. The adaptation of descriptors from this study for a self-assessment pack for students starting a course of study at the university of Basel. Descriptors are presented on a questionnaire in a succession of "mini-scales" for different categories. The aim is to profile the self-assessed proficiency of the students in order to give feedback to the upper secondary sector. Of the 51 descriptors included on the questionnaire, 27 were edited versions of descriptors produced in this study. The scale values of those 27 descriptors remained relatively stable in the new context. The correlation between the original logit values and those produced in an independent analysis was 0.899.

c. A formal assessment of educational achievement in French and English at the end of primary school and lower secondary school in central Switzerland planned for 2001. Here, adapted Portfolio checklists of descriptors will be used for teacher and self-assessment. They will be supplemented by communicative assessment tasks operationalising a sample of the descriptors and conventional tests. Rating schemes for oral assessment will link locally relevant, task-related rating scales to the framework of the Common Reference Levels.

A second area for follow up would be to see whether the scale values produced and replicated in Switzerland can be replicated elsewhere. It would be interesting to conduct a survey outside Switzerland to see to what extent the proficiency of school and general adult learners in neighbouring countries is described by the scale. One might hazard a guess that the situation for Francophone learners in a French pedagogic culture in France might not be that different from their colleagues in the Romandie. The same may be true with German-speaking Switzerland and Germany where the school systems and pedagogic cultures are similar. What will happen with learners further afield is difficult to say.

Results from the DIALANG project set up by the European Union are certainly encouraging. In DIALANG descriptors taken from this study are being used for self-assessment at the start of a computer-adaptive test. Initial trialling in 1999 involved 304 learners of Finnish, of whom 254 were

Swedish speakers using the descriptors translated into their mother tongue. Only descriptors for Listening, Reading and Writing were included. Those descriptors for Listening and Reading originated mostly from the 1995 Swiss follow up survey, equated to the original 1994 Interaction-Spoken Production scale. Those descriptors for Written Production were given the scale values of the original descriptors scaled in relation to Spoken Production. Despite the fact that the three skills involved in the DIALANG study were *not* among those scaled in the 1994 survey which produced the common scale, the correlation between the original logit values and those from the DIALANG analysis was again 0.899.

A third follow up is take the descriptor-bank concept literally and establish links between this bank of descriptors and another bank in order to equate the two scales. UCLES are doing this at the time of writing (early 2000) whilst calibrating the ALTE "Can do statements." A set of 16 descriptors based on the sub-scale for Fluency and other "Excellent Items" concerned with a broader view of communicative fluency were distributed through the booklets used for data collection in order to serve as anchor items to link the two data sets. The correlation here for the scale values of these 16 anchor items in the original 1994 analysis and in the 2000 ALTE analysis was 0.97 (n = circa 1,500).

A final, and most significant follow up would be to try and formulate descriptors to describe the areas this study failed to address adequately. The major area of concern is socio-cultural competence. Socio-cultural competence appears to remain an area which needs to be scaled, if it can be scaled, entirely separately from language proficiency.

Other possible areas for future research concern the technical problems with the Rasch model which were encountered:

- The Rasch model appears to severely distort the logit estimates for learners / items with very high or very low scores. The question is how high is high? How low is low?
- The rating scale model (RSM), as operationalised in e.g. FACETS, apparently produces a longer logit scale than does the dichotomous model, as operationalised in e.g. Itembanker. Does this longer scale reflect the reality of more sophisticated judgements or is the RSM chasing its own tail?
- RSM analysis programs appear to ratchet the separate data collection forms too closely together when linking onto a common scale is done automatically? Does this also happen with the dichotomous model?
- Teachers appear to tend to claim too wide a range on the rating instrument. They seem to tend to use criterion statements to separate learners rather than matching learners to the criteria. Is this inevitable, a question of disposition ("sensitivity") or can it be improved through training?
- How do these various factors really interrelate?

These are not necessarily areas for applied linguistics, but they are areas which perhaps have a particular relevance to applied linguistics in view of the increasing trend towards the use of the Rasch model in language testing and in particular in relation to the validation of scales of language proficiency (e.g. Brown et al 1992; Milanovic et al 1992/6; Stansfield and Kenyon 1992; Fulcher 1993; Hamilton et al 1993; Lumley 1993; Lee 1993; Lee et al 1998; current ALTE work).

Nevertheless, forewarned is forearmed. The experience in this study would suggest that corrections can be applied for all these problems. The stability of the scale values in the 1995 follow up, the Basel university

project, in the DIALANG and ALTE validation studies suggests that the corrective devices adopted in this study did work. The Rasch model in general and the program FACETS in particular is an extremely useful tool for putting teacher judgements into a measurement framework, for investigating the communality of that framework, for investigating the number of decision strata (“levels”) in the data, for placing learners as well as items onto a common scale, and for identifying the need for and effects of rater training. It is thus uniquely relevant to the development of a common framework.

The Development of a Common Framework Scale of Language Proficiency