Publié dans Bulletin VALS-ASLA (Association suisse de linguistique appliquée) 64, 129-143, 1996
qui doit être utilisée pour toute référence à ce travail

Description and Assessment of Foreign Language Learning Proficiency in the Swiss Educational System

Brian NORTH

Abstract

Kompetenzskalen zur Beurteilung und Beschreibung des fremdsprachlichen Könnens beruhen in der Regel auf der subjektiven Erfahrung und dem Konsens ihrer Verfasser. Im Projekt "Evaluation und Selbstevaluation der Fremdsprachenkompetenz an Schnittstellen des schweizerischen Bildungssystems" wurde eine solche Kompetenzskala auf empirischem Weg entwickelt. Der Artikel beschreibt und diskutiert die Methoden und die vorläufigen Ergebnisse des Projekts. Kurze transparente Beschreibungen für das, was Lernende auf verschiedenen Niveaus mit der Fremdsprache tun können, wurden in Workshops auf ihre Qualität hin überprüft und von Fremdsprachenlehrern für Englisch, Deutsch und Französisch zur Beurteilung von Lernenden verwendet. Für die fremdsprachliche Kommunikationsfähigkeit in der Interaktion und für das Hörverstehen in Situationen der Einwegkommunikation konnte mit Hilfe einer Rasch-Analyse eine gemeinsame zehnstufige Skala abgeleitet werden, die eine Übersicht über die von Lernenden in verschiedenen Bildungsinstitutionen und Sprachregionen erreichten Niveaus ermöglicht. In der Schlussphase des Projekts wird die Nutzung der Kompetenzbeschreibungen und -skalen für die Selbstevaluation und für die Entwicklung eines vom Europarat geplanten Fremdsprachenportfolios geprüft.

1. Context

This paper summarises provisional results from the project *Evaluation et auto-évaluation de la compétence en langues étrangères aux points d'intersection du systeme d'enseignement suisse* (SCHNEIDER & RICHTERICH) which aims to create a "Swiss Framework" of levels defined with transparent descriptors of what learners are capable of doing at each level and to develop prototypes for a "Language Passport" or "Language Portfolio" to record achievement in relation to an internationally recognised framework of reference. The project is a follow up to the Council of Europe Intergovernmental Symposium "*Transparency and Coherence in Language Learning in Europe: objectives, evaluation and certification*" (COUNCIL OF EUROPE 1992) and the scaled bank of descriptors of communicative language proficiency developed during the project has been used to define the levels and illustrate some of the categories proposed in draft 1 of the Council of Europe proposal for a Common European Framework of reference for language learning and teaching (COUNCIL OF EUROPE 1996).

The project bases itself on two models: (a) the descriptive model of language proficiency put forward in the Council of Europe Framework which is related by NORTH (1994) to existing models of competence and proficiency (e.g. CANALE & SWAIN 1980; VAN EK 1986; BACHMAN 1990); (b) the Rasch Item Response Theory measurement model (WRIGHT & STONE 1979; WOODS & BAKER 1985) in a scalar variant which takes account of assessor subjectivity (LINACRE 1989). The methodology adopted is a development of that proposed by NORTH (1993a) and the survey of English conducted to pilot the methodology in 1994 has been the subject of a PhD thesis (NORTH 1995).

In the implementation phase just beginning, teachers in the network created by the project are being invited to experiment with the descriptors in continuous assessment instruments for teacher and learner use: for example *checklists* (e.g. OSCARSON 1978) and *profile grids* (c.f. BRINDLEY 1989 for a review of formats).

2. Scaling Proficiency Descriptors

The creation of a transparent, coherent framework of defined levels and categories presupposes assigning descriptions of language proficiency to one level or another - that is *scaling* descriptors. Considering the extent of the literature on scaling and on behavioural scaling in particular (e.g. SMITH & KENDALL 1963; LANDY & FARR 1983; BORMAN 1986), it is in fact surprising how little use has been made of scaling theory or of empirical development in the production of language proficiency *scales*, virtually all of which appear to have been produced on the basis of intuition by a single author or small committee (c.f. NORTH 1993b for reviews). Yet subjectivity in assessment with reference to defined criteria occurs in two quite separate ways. Firstly and most obviously, raters vary widely in their severity (CASON & CASON 1984; LINACRE 1989) which is why assessment approaches involving two assessors are increasingly common. Secondly, however, the assignment of a descriptor to a level by the author(s) systematises subjective error on the part of the author(s) so that even if raters are trained to assess the same way (to reduce the difference in severity between them) the validity of the assessment is still questionable - and reliability will be hard to achieve as those descriptors which are poorly defined and/or incorrectly placed on the scale will undermine the assessors efforts. Whilst this approach may work in an in-house assessment approach for a specific context with a familiar population of learners and assessors, it has been criticised in relation to the development of national framework scales of language proficiency (e.g. SKEHAN 1984; FULCHER 1987, 1993 in relation to the British ELTS; BRINDLEY 1986, 1991 in relation to the Australian SECOND LANGUAGE PROFICIENCY RATINGS (ASLPR); BACHMAN & SAVIGNON 1986, LANTOLF & FRAWLEY 1985, 1988; SPOLSKY 1986, 1993 in relation to the American Council on the Teaching of Foreign Languages (ACTFL) Guidelines). The problem is well known in the scaling literature and led THURSTONE (1928:547-8, cited in WRIGHT & MASTERS 1982:15) to propose that:

> «the scale values of the statements should not be affected by the opinions of the people who helped to construct it (the scale). This may turn out to be a severe test in practice, but the scaling method must stand such a test before it can be accepted as being more than a description of the people who construct the scale. At any rate, to the extent that the present method of scale construction is affected by the opinions of the readers who help sort out the original statements into a scale, to that extent the validity of the scale may be challenged.»

For the development of a scale which is intended to serve as a point of reference for different educational sectors, linguistic areas and target languages, these problems are particularly acute. No small group of "readers" would be sufficiently representative to arrive at generalisable scale values, no informants could provide information about the way in which descriptors perform when actually used to assess the learners in question. A Rasch analysis calibrates the items (here descriptors of proficiency) and the persons (here language learners) onto the same arithmetical scale and in addition offers the opportunity to identify those items or people who do not "fit" with the main construct being measured and exclude them if desired in order to increase the accuracy of the calibration of the difficulty values for those descriptors which are consistent and reliable and the calibration of the learner proficiency values from those teacher ratings which are consistent and credible. Those proficiency values can be adjusted to take account of the degree of severity/lenience of the assessor and the result obtained gives probably the nearest thing to objective measurement of subjective judgements which is currently technically feasible.

3. Project Phases

The overall structure of the project is given in the chart below:

Year 1	English	Spoken & Written Interaction Spoken Production (Monologue)	Teacher Assessment
Year 2	French German English	Spoken Interaction & Production Receptive Listening Reading	Teacher & Self Assessment

The focus in the pilot for English in Year 1 was on Spoken Interaction, including Comprehension in Interaction. Descriptors were also included for Spoken Production (extended monologue: describing, putting a case) and for Written Interaction (letters, notes, form-filling).

In Year 2, the survey was extended to French and German. Approximately a third of the 212 descriptors calibrated in Year 1 were reused in order to link the two surveys and descriptors were added for Reception: for Reading and for non-interactive Listening.

The project followed a broadly similar pattern of three phases each year: (a) the creation of a descriptor pool; (b) qualitative validation in workshops with teachers, and (c) quantitative validation of checklist assessment of learners.

3.1. Creating a Descriptor Pool

A comprehensive review of existing scales of language proficiency undertaken for the Council of Europe (NORTH 1993b) provided a starting point. Definitions from different scales were assigned to provisional levels and each definition was then split up into sentences, with each sentence being allocated to a provisional category. Where possible the categories were related to those emerging in the work of the Council of Europe Framework authoring group for (a) Communicative activities, (b) Strategies; (c) Aspects of communicative language proficiency. There were virtually no descriptors for Strategies so about 80 new descriptors were written and statements about qualitative aspects of writing which might equally apply to speaking were also amalgamated into the pool. With the elimination of straight repetition, negative formulation and norm-referenced statements now meaningless away from their co-text a pool of approximately 1,000 descriptors was developed in each of the two years. These were edited where necessary to create stand-alone, positively worded criterion statements.

3.2. Qualitative Validation: Workshops with Teachers

Having constructed a pool of likely descriptors, the next step was to find out which of them (a) described what they seemed to describe (b) were relevant to the kinds of learners concerned (c) covered the things teachers wanted to say and (d) were interpreted consistently with regard to approximate level. The confirmation of these points was the aim of a series of 32 workshops, which followed a similar pattern both years.

In the first technique, an adaptation of one used by POLLITT & MURRAY (1993), teachers were asked to discuss which of a pair of learners talking to each other on a video was better - and justify their choice. The aim was to elicit the

metalanguage teachers used to talk about qualitative aspects of proficiency and check that these were included in the categories in the descriptor pool.

The second technique was a sorting task based on that used by SMITH & KENDALL (1963) in the development of arguably the first defined assessment scale with calibrated descriptors. Pairs of teachers were given a pile of 60-90 descriptors cut up into confetti-like strips of paper and asked to sort them into 3-4 labelled piles, which represented potential categories of description. The categories would be related - for example Fluency, Flexibility, Coherence (all Pragmatic Competence). At least two, generally four and up to ten pairs of teachers (either at the same or succeeding workshops) sorted each set of descriptors. Labels were written on envelopes into which the descriptors were to be put. An extra envelope marked "Unclear / Unhelpful" was also provided for descriptors for which the teachers couldn't decide the category, or found unclear, verbose or otherwise unhelpful. Teachers were also asked to tick those descriptors which they found particularly clear and useful, and (at some workshops) to identify which were relevant to their particular sector and which were suitable for self assessment. Results were recorded as codes which were used to identify clear, relevant, positive descriptors to include in the questionnaire survey.

A final technique checked the consistency with which descriptors were assigned to levels. In an adaptation of THURSTONE's (1928) sorting task at least two pairs of teachers at later workshops in both years were asked to put the surviving descriptors for a particular category into three piles, low - middle - high, and then, when feasible, into two subdivisions of each of the three broad level to give 6 bands. Again results were coded and descriptors which were not interpreted consistently were rejected from the descriptor pool. This technique was also used extensively in Year 2 to identify translation difficulties.

3.3. Quantitative Validation: Checklist Assessment of Learners

A selection of the surviving descriptors was then used to construct a series of questionnaires at different levels. In Year 1, seven questionnaires were used whereas in Year 2 the same range of level was covered with four - with a fifth very high level questionnaire which in the event did not yield enough data for a satisfactory analysis. Each questionnaire consisted of 50 descriptors grouped under appropriate headings, and each descriptor had a 0-4 rating scale which was defined on the cover page of the questionnaire as follows:

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.**
Speaking: for example: if he/she has some time to think about what to say, or the interlocutor is tolerant and prepared to help out.
Listening: for example: if the reception is very clear and/or he/she has a chance to hear it twice and/or can ask occasionally what something means.
Reading: for example: if he/she has the time to reread and/or to consult reference sources and/or can ask for occasional help

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.
Speaking: for example when in a surprising situation or when talking to a less co-operative interlocutor.
Listening: for example when there is an element of audial interference, and/or when speech is rapid and/or when he/she can only hear it once.
Reading: for example: when he/she has only time to read quickly and/or has little chance to study difficult sections.

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

On the pages with the descriptors, the above definitions appeared in short form, as below:

Please cross the appropriate number next to each item: X

0	1	2	3	4
Describes a level beyond his/her capabilities	Yes, in favourable circumstances	Yes, in normal circumstances	Yes, even in difficult circumstances	Clearly better than this

SPOKEN TASKS					
1. Can communicate in simple and routine tasks requiring a simple and direct exchange of information.	0	1	2	3	4
2. Can ask for and provide everyday goods and services.	0	1	2	3	4
3. Can make simple purchases by stating what is wanted and asking the price.	0	1	2	3	4

QUALITIES OF SPOKEN PERFORMANCE					
40 Can make him/herself understood in short contributions, even though pauses, false starts and reformation are very evident.	0	1	2	3	4
41 Can communicate with memorised phrases, groups of a few words and single expressions and formulae.	0	1	2	3	4
42 Can use some simple structures correctly but still makes basic mistakes.	0	1	2	3	4

The methodology used in the analysis was an adaptation of classic item banking methodology in which a series of tests (here questionnaires) are linked by common items called "anchor items". In addition in Year 2, 70 of the 170 items

employed "anchored" back to the 1994 English survey. Linking the questionnaires together in this way was sufficient to be able to relate the descriptors to each other to create a common scale, but in order to calibrate learners onto the scale with teacher ratings, it was necessary to link the teachers together in order to be able to take the variation in their strictness/leniency into account in assessing the learner proficiency (LINACRE 1989). This linking was achieved by asking participating teachers to rate performance on video recordings of learners in the questionnaire survey with 12 appropriate descriptors from the relevant questionnaire. Finally, in Year 2 a Self Assessment questionnaire of 20 descriptors was selected from the questionnaire at each level, with again anchoring through common items.

The adaptation of itembanking methodology to teacher ratings across the full range of proficiency through analysis of forms linked by anchor items was not without its problems. Space does not permit adequate discussion of them in this short paper, but there were two serious ones. Firstly, Rasch model analysis whilst very reliable within what WARM (1989:442) describes as "*rational bounds*" produces distorted values for scores at the two extremes. This problem was anticipated; no items were lost thanks to the pre-testing, but a fair number of learners were removed from the analysis for this reason. Secondly, many teachers showed unmistakable evidence of using the descriptors (i.e. criteria) provided to separate out their learners (if I gave her a "2" I'd better give him a "1") so that they grossly overestimated the range of level in the class. This had little or no effect on the calibration of the descriptors and hence creation of the scale but posed severe problems for determining proficiency values for the learners. A technique based on the standard deviation of the ranges of levels spanned by the ratings for each class was used to identify and correct for the extent to which teachers were norm-referencing in this way.

4. Results

The data analysis has three products (a) a scale of 10 defined levels; (b) a bank of classified descriptors covering a relatively large number of categories related to the Council of Europe Framework, and (c) a map of the achievement of Swiss foreign language learners for English, French and German relating proficiency achieved in different educational sectors to years of study.

4.1. A Common Scale

The descriptors are calibrated in rank order on an arithmetic scale. Levels are created by establishing "cut-off points" on the scale. Setting cut-offs is always a subjective decision. As JAEGER (1976:2; 1989:492) says "*no amount of data collection, data analysis and model building can replace the ultimate judgmental act of deciding which performances are meritorious or acceptable and which are unacceptable or inadequate*" or as WRIGHT & GROSSE (1993:316) put it : "*No measuring system can decide for us at what point "short" becomes "tall"*".

The decision to report 10 levels is, however, far from arbitrary. Firstly as POLLITT (1991:90) shows there is a relationship between the reliability of a set of data and the number of levels it will bear, secondly these cut-offs are set at almost exactly equal intervals on the measurement scale, and finally a comparative analysis of the calibration of the content elements which appear in descriptors (e.g. topics you can handle; degree of help required) and detailed consideration of the formulation of adjacent descriptors shows a remarkable degree of coherence inside these levels - and a *qualitative* change at these cut-off points.

The scale of 10 levels was produced in the 1994 analysis and one of the functions of the 1995 survey was to replicate the 1994 finding. To achieve this the 1995 data was analysed both with the 70 descriptors from 1994 anchored to the difficulty values established in 1994, and entirely separately. A few of these descriptors were shown to be unstable as was the case with the anchoring between questionnaire forms in both years. Instability in a couple of anchors is a relatively common occurrence in itembanking with test data; such items are removed from the analysis as they distort the result.

The Interaction, Listening and Reading items were analysed both together and separately and a large number of analyses were undertaken to look at variation in difficulty values across target languages and demographic variables like language region, educational sector and mother tongue and identify which descriptors kept the most stable values across contexts. Such very stable descriptors would be most suitable for use in the construction of an overall "global scale" for a "Language Passport".

A substantial degree of variation was discovered but there has not yet been time to ascertain its significance since (a) the different variables (e.g. mother tongue, language region) interact and what appears to be an effect of one variable can in fact be a disguised effect of another, and (b) the small sample of teachers involved for most variables on any one questionnaire means that the picture obtained in the majority of sub-analyses could be unrepresentative. Larger scale comparisons are therefore more effective - for example comparing difficulty values arrived at by analysing the ratings from people teaching their own mother tongue (approximately 30% of the total) in contrast to ratings from

people teaching what for them is a foreign language. Here the only items to show statistically significant variation at the 5% level are all Listening or Reading items, that is to say that the (dominant) Speaking construct is totally stable. At the level identified as Threshold Level there is variation only in relation to one unsuccessful item about listening to announcements - later dropped. At a level above Threshold (upper intermediate) there is a slight tendency for native speaker teachers to rate listening items as more difficult than the non-native teachers. At advanced there is a similar effect, this time with reading items: native speaker and non-native speaker teachers appear to mean something slightly different by "understand" in relation to literature and other complex text.

As a result of all this investigation Reading was removed from the main analysis because of the suspicion that the Reading items were behaving differently. When created separately, the scale for Reading was significantly shorter than the scale for Listening & Speaking, which meant that the difficulty of the higher descriptors would have been underestimated had it been kept in the main analysis rather than analysed separately.

The main Listening & Speaking scale, however, appeared very stable. After removing 8 of the 70 descriptors anchoring to 1994 because they were showing significant instability, the values of the 108 Listening & Speaking items from 1995 when (a) analysed alone and (b) analysed with 62 items anchored to 1994 values correlated 0.992 (Pearson). This is a very high consistency between the two years when one considers that (a) 1994 values were based on 100 English teachers, whilst in 1995 only 50 of the 184 teachers taught English so the ratings dominating the 1995 construct were those of the French and German teachers, and (b) the questionnaire forms were completely different (4 in 1995 covering the ground of 7 in 1994).

4.2. A Classified Descriptor Bank

Not all of the categories originally included could be successfully calibrated. Sometimes this was due to a lack of what in the Rasch literature is referred to as "unidimensionality". This has a technical meaning related to the technical meaning of reliability as separability: is it possible to separate out items by their difficulty along the same dimension (a 45^{o} angle on a plot). Are the items strung out nicely along the 45^{o}? Or are some items pulling away to the sides because they do not really "fit" the main construct created by the data? Removal of such "outliers" clarifies the picture and increases the scale length and the reliability and the precision of the difficulty values for the items - in this case descriptors. Unlike classical test theory (CTT: reliability theory) item response theory (IRT: Rasch) does not say such outliers are bad items - but rather that they don't belong here and should perhaps be analysed separately to see if they build their own construct. Thus, as mentioned above, Reading did not appear to "fit" a construct dominated by the overlapping concepts of Speaking and Interaction and needed to be analysed separately. In addition, three groups of categories were actually lost:

1. Socio-cultural competence. It is not clear how much this problem was caused by the concept being separate from language proficiency (and hence not "fitting"), by rather vague descriptors identified as problematic in the workshops, or by inconsistent responses by the teachers.
2. Those descriptors relating to interlocutor factors (need for simplification; need to get repetition/clarification) which are implicitly negative concepts. These aspects worked better as provisos at the end of positive statements focusing on what the learner could do (e.g.: *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.*)
3. Those asking teachers to guess about activities (generally work-related) beyond their direct experience: Telephoning; Attending Formal Meetings. Giving Formal Presentations; Writing Reports & Essays; Formal Correspondence. This could be a problem of dimensionality (as with Reading and Socio-cultural competence) or it could be that teachers were just unable to give the information.

However, the categories for which descriptors were successfully validated and calibrated offer a relatively rich metalanguage to describe proficiency:

Communicative Activities		
Global Language Use: Overall Interaction		
Listening:		Overall Listening Comprehension
	Receptive:	Listening to Announcements & Instructions
		Listening as a Member of an Audience
		Listening to Radio & Audio Recordings
		Watching TV & Film
	Interactive:	Comprehension in Spoken Interaction
Reading:		*Not yet finalised*
Interaction: Transactional:		Service Encounters & Negotiations
		Information Exchange
		Interviewing & Being Interviewed
		Notes, Messages & Forms
Interaction: Interpersonal:		Conversation
		Discussion
		Personal Correspondence
Production (Spoken): (Sustained Monologue)		Description
		Putting a Case
Strategies		
Interaction Strategies:		Turntaking
		Cooperating
		Asking for Clarification
Production Strategies:		Planning
		Compensating
		Repairing & Monitoring
Aspects of Communicative Language Proficiency		
Pragmatic: (Language Use)		Fluency
		Flexibility
		Coherence
		Thematic Development
		Precision
Linguistic: (Language Resources)	Range: (Knowledge):	General Range
		Vocabulary Range
	Accuracy : (Control)	Grammatical Accuracy
		Vocabulary Control
		Pronunciation

4.3. A Map of Learner Achievement

Because learners and descriptors are calibrated onto the same scale and this scale remained the same each year (since the scales for 1994 and 1995 correlate 0.992) it is possible to relate learner achievement to the set of 10 levels identified. There are complications caused by the marked tendency to exaggerate the range of level in a class, but the mean and standard deviation of these ranges was again virtually identical in both the two years, so the corrective action taken was at least consistent. Consideration of the results is not yet complete - there is a degree of examination information which needs to be related to the survey data - but the picture which emerges for English in both years is remarkably coherent with progress clearly related to years of exposure for all sectors. For French the picture appears slightly less clear, with a suggestion that in Gymnasium, although the mean achievement increases with each year, a substantial minority appears to make little or no progress with increased years of study. This same pattern appears even more clearly for German in relation to the adult and lower secondary as well as gymnasium sectors. These findings are, however, still provisional. A certain number of technical problems caused by response effects, data collection design and known weaknesses of the analysis method - all in relation to the measurment of high scoring and low scoring subjects - means that results for the class averages will have a considerably higher reliability than the picture of the full range of achievement for each class discussed above.

The provisional results for self assessment are disappointing. Because the rating scale given to learners was restricted to the first four points (0-3) the self assessment data had to be analysed separately and the teacher assessment and self assessment scales equated through percentiles, which may have complicated matters. Secondly, the removal of the Reading items and items dropped as misfitting or unstable during the course of the analysis reduced the original 20 self assessment items per questionnaire to 11 or 12, - leading to high standard errors equivalent to at least half a level on the 10 level scale. Nevertheless, the correlation for the 208 learners whose teacher and self ratings could be compared was only 0.386 (Pearson) on scores on the common arithmetic scale or 0.392 (Spearman) and 0.413 (Pearson) on the 10 identified levels. Such a level of correlation does not compare that favourably with correlations between self assessment and test scores or teacher assessments of around 0.5-0.7 reported in the literature (see e.g. OSCARSON 1984 for a review). Again, this result is provisional, and does not at the end of the day say who was right! The tendency to exaggerated teacher norm-referencing and technical difficulties in the calibration of the learners has already been referred to, and the results are currently in the process of being compared to examination data, which may contribute to a reinterpretation of the respective assessments.

5. Current Developments

However, this result may well influence the formats adopted for the Language Portfolio. Initial prototypes are currently (May 1996) being circulated to the teacher network for their reactions and include scale, grid and checklist formats. That transparent, coherent descriptors can - in an appropriate format - achieve respectable correlation to teacher ratings moderated for subjectivity is suggested by the results of a Eurocentres study for BIGA relating to 108 long-term unemployed young people sent on Eurocentres courses in 1993. Before the course, these learners were asked to rate themselves on the Eurocentres global scale (which has a short two to three sentence paragraph per level) and their ratings correlated 0.74 (0.78 for English alone) to placement on the Eurocentres scale averaged from an interview and a test from an item bank.

There is an argument that the process of diagnosis and profiling in relation to categories with more specific checklists *(diagnosis-oriented)* should be separated from proficiency assessment in relation to simple, straightforward rating instruments *(assessor-oriented)* (See e.g. HULSTJIN 1985:280; MATTHEWS 1990; POLLITT & MURRAY 1993). This distinction is a principle in the design of the Portfolio: simpler, more holistic and very stable descriptors in the "Passport" for proficiency assessment and recording of qualifications; more detailed information in a "Map" for orienting and recording learning.

The research project has focused on individual descriptors because the aim was to calibrate those descriptors as objectively as possible as stand-alone criterion statements, since it is a criticism of many scales of language proficiency that descriptors have meaning only relative to their textual context - the wording of descriptors above and below them, other statements in the same paragraph - rather than functioning as criterion statements in their own right. One must be careful to avoid confusing means and ends. An appropriate format for data collection is not necessarily an appropriate or valid format for exploitation of the results in assessment instruments. Nor has the validity of the scale produced yet been established when *used* as a scale. It is to these questions of format, exploitation and adaptation, a postieri validation and implementation to which the project in the third and final phase is now turning.

References

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

BACHMAN, L. AND SAVIGNON S.J. (1986): "The Evaluation of Communicative Language Proficiency: A Critique of the ACTFL Oral Interview". *Modern Language Journal*; 70/4:380-90 Win 1986.

BORMAN,W.C. (1986): "Behaviour-based Rating Scales". In BERK, R. (ed) *Performance Assessment: Methods and Applications*, Baltimore MD. The Johns Hopkins University Press.

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

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

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

CANALE, M. & SWAIN, M. (1980): "Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing". *Applied Linguistics* 1/1.1-47.

CASON, G.J.& CASON, C.J. (1984): "A Deterministist Theory of Clinical Performance Rating". *Evaluation and the Health Professions*, 7. 221-247.

COUNCIL OF EUROPE (1992): *Transparency and Coherence in Language Learning in Europe: Objectives, assessment and certification*, Strasbourg, Council of Europe; the proceedings of the Intergovernmental Symposium held at Rüschlikon November 1991 (ed North, B.).

COUNCIL OF EUROPE (1996): A *Common European Framework for Language Learning and Teaching: Draft 1 of a Framework Proposal*, CC-LANG (95) rev III, Strasbourg, Council of Europe.

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

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

HULSTJIN, J.H. (1985): "Testing Second Language Proficiency with Direct Procedures. A response to Ingram". In HYLTENSSTAM, K. and PIENEMANN, M. (eds), *Modelling and Assessing Second Language Development*, Multilingual Matters.

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

JAEGER, R.M. (1989): "Certification of Student Competenc"e. In LINN R.L. (ed) *Educational Measurement*, 3rd edition American Council on Education/Macmillan, New York.

LANDY, F.J. & FARR, J. (1983): *The Measurement of Work Performance*, San Diego, CA. Academic Press.

LANTOLF, J. AND FRAWLEY, W. (1985): "Oral Proficiency Testing: A Critical Analysis". *Modern Language Journal* 70: 337-345.

LANTOLF, J. AND FRAWLEY, W. (1988): "Proficiency, Understanding the Construct". *Studies in Second Language Acquisition* 10/2: 181-196.

LINACRE, J.J. (1989): *Multi-faceted Measurement*, Chicago, MESA Press.

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

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

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

NORTH, B. (1994): *Perspectives on Language Proficiency and Aspects of Competence: A Reference Paper discussing issues in defining categories and levels*, Council of Europe CC-LANG (94) 20.

NORTH, B. (1995): *The Development of a Common Framework Scale of Descriptors of Language Proficiency Based on a Theory of Measurement*, Unpublished PhD thesis, Thames Valley University.

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

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

POLLITT, A. (1991): "Response to Alderson, Bands and Scores". In ALDERSON J.C & NORTH B. (eds), *Language Testing in the 1990s*. Modern English Publications/British Council, Macmillan: 87-94.

POLLITT, A. AND MURRAY, N.L. (1993): What Raters Really Pay Attention to. Paper presented at the 15th Language Testing Research Colloquium, Cambridge and Arnhem, 2-4 August 1993.

SKEHAN, P. (1984): "Issues in the Testing of English for Specific Purposes". *Language Testing* 1(2), 202-220.

SMITH, P.C.AND KENDALL, J.M. (1963): "Retranslation of Expectations: An Approach to the Construction of Unambiguous Anchors for Rating Scales". *Journal of Applied Psychology*, Vol 47/2.

SPOLSKY, B. (1986): "A Multiple Choice for Language Testers". *Language Testing* 3/2: 147-158.

SPOLSKY, B. (1993): "Testing and Examinations in a National Foreign Language Policy". In SAJAVAARA, K., TAKALA, S., LAMBERT, D. and MORFIT, C. (eds) *National Foreign Language Policies: Practices and Prospects*. Institute for Education Research, University of Jyväskyla: 194-214.

THURSTONE, L.L. (1928): "Attitudes Can be Measured". *American Journal of Sociology*, 33 529-554; cited in WRIGHT, B.D. & MASTERS, G. (1982) *Rating Scale Analysis: Rasch Measurement*, Chicago, Mesa Press: 10-15.

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

WARM, T.A. (1989): "Weighted Likelihood Estimation of Ability in Item Response Theory", *Psychometrika* 54:3, Sep 1989: 427-450.

WOODS, A. AND BAKER, R. (1985): "Item Response Theory". *Language Testing* 2.

WRIGHT, B.D. & GROSSE, M. (1993): "How to Set Standard"s, *Rasch Measurement*, Transactions of the Rasch Measurement Special Interest Group of the American Educational Research Association, 7/3 Autumn 1993: 315-6.

WRIGHT, B. D.& STONE, M.H. (1979): *Best Test Design*, Chicago, Mesa Press.