ED 365 094 FL 021 559

AUTHOR North, Brian

TITLE The Development of Descriptors on Scales of Language Proficiency. NFLC Occasional Papers.

INSTITUTION Johns Hopkins Univ., Washington, DC. National Foreign Language Center.

PUB DATE Apr 93

NOTE 71p.

AVAILABLE FROM NFLC, Johns Hopkins University, 1619 Massachusetts Ave., N.W., Washington, DC 20036 ($5 prepaid).

PUB TYPE Information Analyses (070)

EDRS PRICE MF01/PC03 Plus Postage.

DESCRIPTORS Classification; Communicative Competence (Languages); *Evaluation Criteria; *Item Response Theory; *Language Proficiency; Measurement Techniques; Models; *Rating Scales; Scoring; Standards; *Testing Problems; Test Use

IDENTIFIERS *ACTFL Proficiency Guidelines; *Rasch Model

ABSTRACT

Theoretical issues underlying the development of scales of language proficiency are examined. First, a brief classification of scale types is presented, and problems identified in them and advantages they offer are outlined. Discussion then moves to the issues of describing and measuring language proficiency. In this section, behavi ally-based scales in work performance, the historical development of the American Council on the Teaching of Foreign Languages/Interagency Language Roundtable (ACTFL/ILR) proficiency guidelines, the utility of the Rasch model in rating framework development, and subjectivity in judgments and observed behavior ratings are discussed. A many-faceted version of the Rasch item-response theory model is suggested as an instrument that may aid in the mapping of behaviors onto a common metric while identifying different interpretation of the same descriptors by different types and groups of users. Contains about 250 references. (MSE)

The Development of Descriptors on Scales of Language Proficiency

BRIAN NORTH
Eurocentres Foundation

The National Foreign Language Center
Washington DC

About the Occasional Papers

This is one in a series of Occasional Papers published by the National Foreign Language Center. The NFLC prints and distributes articles on a wide variety of topics related to foreign language research, education, and policy. The Occasional Papers are intended to serve as a vehicle of communication and a stimulant for discussion among groups concerned with these topic areas. The views expressed in these papers are the responsibility of the authors and do not necessarily reflect the views of the NFLC or of the Johns Hopkins University.

About the National Foreign Language Center

The National Foreign Language Center, a nonprofit organization established within the Johns Hopkins University in 1987 with support from major private foundations, is dedicated to improving the foreign language competency of Americans. The NFLC emphasizes the formulation of public policy to make our language teaching systems responsive to national needs. ts primary tools in carrying out this objective are:

- *Surveys.* The NFLC conducts surveys to collect previously unavailable information on issues concerning national strength and productivity in foreign language instruction, and our foreign language needs in the service of the economic, diplomatic, and security interests of the nation.
- *National policy planning groups.* In order to address major foreign language policy issues, the NFLC convenes national planning groups that bring together users of foreign language services and representatives of the language instructional delivery systems in formal education, the government, and the for-profit sector.
- *Research.* The NFLC conducts research on innovative, primarily individual-oriented strategies of language learning to meet the nation's foreign language needs of the future.

In addition, the NFLC maintains an Institute of Advanced Studies where individual scholars work on projects of their own choosing.

The results of these surveys, discussions, and research are made available through the NFLC's publications, such as these Occasional Papers, and they form the basis of fresh policy recommendations addressed to national leaders and decision makers.

Price per copy: $5.00, prepaid

The Development of Descriptors on Scales of Language Proficiency: Perspectives, Problems, and a Possible Methodology Based on a Theory of Measurement

Brian North
Eurocentres Foundation

This paper looks at some of the theoretical issues underlying the development of scales of language proficiency, which have become increasingly popular during the last ten to fifteen years. The paper starts by giving a brief classification of types of scales, then lists problems that have been noted in connection with them and the attractions they offer. It then discusses the description issue and the measurement issue in proficiency scale development. A many-faceted version of the Rasch item-response theory model is suggested as an instrument that may aid the mapping of behaviors onto a common metric while identifying different interpretation of the same descriptors by different types and groups of users.

The immediate background to the paper is moves in Switzerland toward the development of a common framework identifying the levels of proficiency achieved by language learners at the points at which they switch between educational sectors. Such a framework is complicated by the decentralization of the Swiss system, in which the twenty-six cantons run their own educational systems, and by the multilingual, multicultural nature of the country. This project is a Swiss contribution to the moves toward a common European framework for

Brian North is a pedagogical adviser with Eurocentres Learning Service, Zurich. Eurocentres—the Foundation for European Language and Educational Centres--is a nonprofit foundation based in Switzerland, with centers offering language courses to adults in countries in which the languages concerned are spoken. Mr. North was previously head of studies of Eurocentre Bournemouth, England, and since 1983 has been involved in the development of scales of proficiency, particularly the use of defined scales to rate group oral interaction in communicative activities, and the development of transparent definitions for user scales. He has been active in the International Association of Teachers of English as a Foreign Language (the British equivalent to Teachers of English to Speakers of Other Languages), and organized the 1989 IATEFL Language Testing Colloquium "Language Testing in the 1990s." He was the coordinator of the Council of Europe Intergovernmental Symposium "Transparency and Coherence in Language Learning in Europe," hosted by the Swiss government in Ruschlikon in 1991.

This paper was written at the end of a three-month stay as a Mellon fellow at the National Foreign Language Center, which provided an opportunity to investigate some of the issues involved in developing or validating scales of proficiency

language learning following the Council of Europe Intergovernmental Symposium "Transparency and Coherence in Language Learning in Europe: Objectives, Assessment, and Certification," hosted by the Swiss government at Rüschlikon, near Zurich, in November 1991.

1. INTRODUCTION

Types of Scales of Language Proficiency

Scales of language proficiency are becoming increasingly popular as a way of creating transparency and coherence in educational systems and subsystems. Although the literature on scales of proficiency is thin on the ground—with the exception of the great debate generated by the Proficiency Guidelines developed by the American Council on the Teaching of Foreign Languages (ACTFL) on the basis of the U.S. government system (ACTFL 1986; Byrnes and Canale 1987)—there are a large number of systems already in regular operation. Ingram and Wylie (1989) present scales in a series of dichotomies, which are glossed and given an example in Table 1.

In discussing types of scales of language proficiency, Alderson distinguishes three types: those that provide quantitative information about tasks, the kind of information that is useful for syllabus and test writers (constructor-oriented); those that provide qualitative information about degrees of skill in the performance (assessor-oriented); and those used to report results to nonspecialists (user-oriented) (Alderson 1991a). North draws attention to the way in which a simplification and synthesis of both constructor- and assessor-oriented (i.e., quantitative and qualitative) information might be appropriate in the descriptors for user-oriented scales in a common reporting framework (North 1992b). In the common framework for English, French, German, Spanish, and Italian that Eurocentres is completing this year with the implementation of a common ten-point scale of language proficiency, these three kinds of information are kept largely separate—though there is a coherent thread running through the different faces of the system. In practice, the original constructor-oriented scales are seldom themselves used; the language specifications that are an analysis of their content have been found more useful for syllabus and materials organization and for test construction. Assessor-oriented assessment scales and criteria subscales have been developed, as have user-oriented certification scales.

This distinction between constructor-oriented and assessor-oriented, quantitative and qualitative, information is similar to the distinction made by Bachman between "real life" approaches like ACTFL and the Australian Second Language Proficiency Ratings (ASLPR), which furnish information about typical tasks a user can do at a level, and what he calls an "ability/interaction" approach, which looks at how the features of the learner's underlying competence interact in the performance of the task (Bachman 1990b, p. 325).

Van Ek has noted a distinction between, on the one hand, a "general characterisation" in scale descriptors designed to ensure "an easy overview of the most relevant points by a wide range of interested parties" and, on the other hand, detailed specifications designed to "guide the planning of learning and assessment activities" (Van Ek 1987). In the case he quotes (Elviri et al. 1986/87) these specifications take the form of sample tests, but in others (e.g., ELTDU, Eurocentres) they take the form of lists of tasks, functions, and structures.

This draws a further significant distinction between scales with language specifications and scales without them. Scales with specifications, of which

Table 1

Whole range of proficiency *Example:* English Speaking Union Framework Project, which calibrated British English as a foreign language (EFL) examinations	**Partial, relevant range** *Example:* English National Curriculum, which covers the range of proficiency of English schoolchildren in ten levels
Serial (a continuous scale) *Example:* English Speaking Union Framework Project, ACTFL	**Threshold** (like a set of exams, which by covering all the relevant levels gives a series of pass/fail thresholds) *Example:* Cambridge EFL exams, which now make up a five-point threshold scale; this scale is being adopted by the Association of Language Testers in Europe
General proficiency *Example:* ACTFL, Eurocentres; these tend to have an overall scale, or scales for the four skills, or both	**Language for specific purposes** *Example:* English Language Teaching Development Unit (ELTDU) and IBM France, each a grid of subscales for specific language activities
Tasks only (quantitative: a list or stringing together of tasks the person can do) *Example:* Early ACTFL	**Total behavior** (quantitative and qualitative: focused on the degree of skill in the performance as well as the tasks) *Example:* Royal Society of Arts, Eurocentres
Proficiency (a holistic concept: development along a continuum) *Example:* All of the above	**Graded objectives** (mastery of a specified list of tasks) *Example:* Scottish Vocational Educational Council, English graded-objectives schemes
Four-skill *Example:* All of the above except ELTDU and IBM	**Overall** *Example:* Finnish Foreign Language Diploma for Professional Purposes
Absolute, noncompensatory (if you do not get passed on specified points, you fail the level) *Example:* Interagency Language Roundtable/ACTFL and exam/threshold models	**Global, holistic** (one synthetic judgment, which level is most applicable) *Example:* All the rest of the above

ELTDU was the first (ELTDU 1975), appear to be inspired directly or indirectly by the Council of Europe specification for the "threshold level" (Van Ek 1975; Coste 1976; Van Ek and Trim 1990). They are primarily concerned with how you get people to the levels described on the scale. Scales without specifications, of which the U.S. FSI (Foreign Service Institute) or later ILR (Interagency Language Roundtable) scale was the first, concentrate on product assessment and are usually primarily concerned with whether or not someone has passed a particular threshold (for a job, to attend a course) or with awarding a diploma.

Problems with Scales of Proficiency

Brindley has listed the problems identified in relation to scales of language proficiency—starting from the fact that it is very difficult to obtain any information about how the descriptors in them were arrived at. His points are summarized and glossed below (Brindley 1991):

1. The logic is circular—the levels equal the criteria, and vice versa. A person is Level 3 because he or she can do these tasks; these tasks are Level 3 because people at Level 3 have been found able to do them (Lantolf and Frawley 1985, 1988, 1992).

2. The incremental shape fails to take into account both backsliding (Pienemann, Johnson, and Brindley 1988) and differential abilities in different domains or "discourse worlds" (Douglas and Selinker 1985; Zuengler 1989). Any assumption in the scale that grammar and phonological ability increase in a linear fashion is contradicted by SLA (second language acquisition) research, which has shown variability to be dependent on such factors as psychosociological orientation (Meisel, Clahens, and Pienemann 1981); emotional investment in the topic (Eisenstein and Starbuck 1989); the discourse demands of the task; the desired degree of convergence/divergence (Rampton 1987); planning time (Ellis 1987); and the ethnicity and status of the interlocutor (Beebe 1983).

3. Descriptors are covertly norm-referenced, and there often seems to be no principled relationship between performance features that appear in one level (Skehan 1984; Brindley 1986).

4. It is very difficult to specify relative degrees of mastery with sufficient precision. As Alderson has asked, "Is 'some' more than 'a few' but fewer than 'several' or 'considerable' or 'many'—and how many is 'many'?" (Alderson 1991a).

5. For a framework identified with an interview system like ACTFL, the range of roles available to the speaker is severely restricted (Lantolf and Frawley 1988; Raffaldini 1988; Van Lier 1989; all acknowledged in Clark and Lett 1988).

6. Descriptors are highly context-dependent, which prevents generalization.

7. Interviews confuse the trait (what they are measuring) with the method used to elicit it. The interviewer is judging a performance in an artificial exchange that he or she is also responsible for (Bachman 1987/88).

8. A lack of upper and lower reference points (perfection, zero) makes a strict application of criterion-referenced theory impossible. One could comment that this would also make any other application of criterion-referencing impossible too, so this objection is perhaps overstated. (The term *criterion-referenced* was apparently first used by Glaser, in an article in which he stated, "Underlying the concept of achievement measurement is the notion of a continuum of knowledge acquisition ranging from no proficiency at all to perfect performance. An individual's achievement level falls at some point on this continuum as indicated by the behaviors he displays during testing." Glaser 1963, p. 519, cited in Glass 1978, p. 240.)

To Brindley's list can be added the following problems:

9. Use of vague generalizations that command acceptance because they can be interpreted to mean different things by different people (Trim 1978).

10. Use of negative, demotivating, norm-referenced wording that fails to recognize the possibility of good performance on low-level tasks and has little relevance to curriculum planning (Trim 1978, 1991).

11. Constructing descriptors in an almost mechanical style, in such a way that the difference between sentences describing a particular underlying competence at two adjacent levels is limited to two words—e.g., the substitution of *difficult* for *moderate* in one half of the sentence and of *good* for *adequate* or *some* for *most* in the other. This reduces readability, particularly for outsiders (North 1992b).

12. Allocating "key tasks" to levels without a principled basis, tapping into convention and cliches among teachers and textbook and scale writers (North 1992b).

13. Confusing cognitive complexity with linguistic complexity in the hierarchy of tasks (Mareschal 1977—though his example of a solution, the ELTDU scale, does precisely that).

14. A tendency in communicative language teaching and testing—and even in SLA research (Lantolf and Frawley 1984)—to equate "communication" with the passing of information. This is perhaps a reaction against the practice in interviewing and teaching whereby the dominant partner asks questions to which he or she already knows the answer, inviting a display-language sample. Terms like *input*, *output*, and *feedback* are examples of how information technology influences our thinking, even though the giving of information is never a primary communicative goal (Sinclair 1985).

15. The common practice of relating scales to a needs analysis, one school of which (Munby 1978) can lead to a reductive, analytical treatment of language to the detriment of the development of "interactional competence" (Kramsch 1986). The actual execution of conversations—the way in which turns are taken, addressees selected, the floor held, and so on (the "move structure")—and the real-time, unpredictable, purposeful nature of spoken communication (Sinclair 1979, 1981) need to be taken into account.

In view of the comprehensiveness of this list of problems, it is perhaps surprising that anyone would attempt to develop a scale of language proficiency. With the notable exception of one or two contributors to the ACTFL debate—for example, Lantolf and Frawley, who appear to consider any form of scale or set of ascending levels or framework unacceptable on philosophical grounds (Lantolf and Frawley 1985, 1988, 1992)—most of those who have identified the problems listed above are people who are interested in the possibilities for transparent and coherent criterion-referenced assessment that scales of proficiency offer.

Attractions of Scales of Proficiency

The main attraction of scales of proficiency made up of defined bands or levels is their ability to provide a unifying framework to

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

— provide guidelines for test construction (Dandonoli 1987; Dandonoli and Henning 1990; Alderson 1991a);

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

— provide coherent internal links in a system among precourse or entry testing, syllabus planning, materials organization, progress and exit assessment, and certification (North 1991);

— permit comparison between systems using a common metric or yardstick (Lowe 1983; Liskin-Gasparro 1984; Bachman and Savignon 1986; Carroll and West 1989); and

— establish a frame of reference that can describe achievement in a complex system in terms meaningful to all the different partners in or users of that system (Trim 1978; Brindley 1986, 1991; Schneider and Richterich 1992).

In addition—provided either that the units of the scale are small enough to measure increments of learning in the school system (North 1992b), or that school grade scales can be designed in relation to broader bands on a proficiency scale (Ingram and Wylie 1989)—scales of proficiency can be used to provide achievement stages and grades that reflect the curriculum of the classroom but that can be translated into a proficiency statement and grade on a common framework. Such achievement grades could be "soft data" steps assessed by informal testing on the way to key points where "hard data" is obtained from examinations. This is the idea behind the Council of Europe European Language Portfolio (Schärer 1992; Council of Europe 1992). Experience with an achievement certification system anchored to known public examinations, as well as defined criteria in the English Eurocentres, tends to suggest that teachers can internalize the standards of a dual-function approach.

Lantolf and Frawley suggest that one should do nothing about framework development until the problem of describing proficiency has been undertaken and a model has been arrived at that everyone can agree with and that has been empirically verified (Lantolf and Frawley 1985, 1988, 1992). Spolsky has voiced a similar concern; however, in his case it appears more a concern that while such systems can operate well in defined domains within a fairly homogeneous institutional group dealing with relatively predictable subjects (e.g., examiners for the U.S. ILR, for Eurocentres, for the Royal Society of Arts/Cambridge Certificate of Communicative Skills in English), the state of knowledge on language proficiency makes the design of a common metric, a common frame of reference, premature (Spolsky 1992). Put another way, this concern characterizes a common framework as an unattainable perfect goal—a holy grail.

In discussing the shortcomings and limitations of scales of proficiency, there is an important distinction to be made between a *theoretical model* to describe the nature of foreign language proficiency, and an *operational model* that people can actually use. An operational model is always simpler than a theoretical model, and while it relates to theoretical models, it may reinterpret elements to make them more accessible in a particular context. Even theoretical models do not describe reality. Rather, they "make ideas about experience explicit. They specify how experience might be simplified so that it can be remembered and managed" (Wright and Masters 1982, p. 60). Even a mathematically rigorous operational scaling model has a very limited operational aim: "to approximate a limited but reproducible continuity" (Wright and Masters 1982, p. 6). In this sense, then, the criticism by Lantolf and Frawley—that scales of proficiency (in this case the ACTFL Guidelines) model reality rather than mirroring it, that they have "constructed a reality" and are "prescriptions of a theorist deciding what speakers ought to do"—is simply misguided (Lantolf and Frawley 1985). All models—including, naturally, all testing models, all syllabus models, all discourse models, all SLA models, and so on—model reality; that is why they are called models.

Lantolf and Frawley criticize the ACTFL Guidelines because they are finely honed, committee-produced, "lovely symmetrical" descriptors (circular logic, point 1 in the list above). The symmetry in the ACTFL system actually goes further than Lantolf and Frawley state; it is indeed surprising the extent to which proficiency is defined in the ACTFL literature as what is tested in the Oral Proficiency Interview—which is defined as the operationalization of the guidelines, which define proficiency (confusing the trait with the methoa: Bachman 1987/88, point 7 in the list above). That is a characteristic, and a weakness, of the ACTFL/ILR system, exacerbated by using grammatical forms as anchoring criteria for particular levels (consistent use of the past tense = advanced). This systematizes two kinds of random error—that of method effect, and that of inappropriate anchors.

Systematizing the random error of method effect. Getting everybody to interview the same way, and to rate the same way (Bachman and Savignon 1986, point 6 above), is a problem with all subjective assessment systems that cannot make an adjustment for task difficulty and rater severity (as is now possible in at least development settings, discussed later in this paper). It is particularly a problem with interviews that are ritualized unequal encounters in which the interviewer is defending counsel, jury, and judge all at once and the dominated partner has a very restricted range of roles (point 5 above: Raffaldini 1988; Kramsch 1986; Shohamy 1988; North 1992a).

Many interview systems try to address this problem by eliciting phases where discourse of radically different sorts takes place—increasingly with two examiners (as in the original FSI/ILR interviews), and increasingly with a mix of native/non-native-speaker and non-native/non-native-speaker talk in an attempt to have "washback validity" (Morrow 1986) or "systemic validity" (Fredericksen and Collins 1989) to help the generation of interactional competence (Kramsch 1986). Byrnes's claim to find discourse competence in the Oral Proficiency Interview (Byrnes 1989) misses the point that discourse competence is inseparable from language. Discourse analysis was developed precisely to find out what was going wrong in unequal encounters (classrooms, doctor-patient interviews, management-union negotiations) and how the structure of the situation needed to be changed so as to improve interactional competence—and successful outcomes (Sinclair 1985).

Systematizing the random error of inappropriate anchors. Everybody has prejudices—personal criteria that make shortcuts in whatever the official system is; the question that "sorts people out"; the rule of thumb that says, "I find that people who can do this are intermediate." The reason for attaching descriptions to levels is to make the criteria applied explicit, shared, and consistent. If the anchors are not chosen on the basis of theory, experience, and empirical item analysis, if a simplistic assumption is used to anchor a key threshold between two levels, error is systematized (Landy and Farr 1983). Using specific linguistic forms as anchors (e.g., consistent use of the past = advanced, as in ACTFL), rather than making a holistic judgment about range and accuracy so as to determine where someone is in his or her development, is a risky business made riskier if it does not take into account what is known from SLA research about fixed and variable sequencing (Pienemann and Johnson 1987). It is this problem that people are referring to when they characterize the ACTFL system as a discrete-point approach, and it is for this reason tnat some systems (e.g., Eurocentres) separate language specifications (constructor-oriented, for input) from scale descriptors (assessor- and/or user-oriented).

To be fair to ACTFL raters, Magnan's study of what actually happens in ACTFL interviews suggests that at least some raters do not actually equate linguistic forms with levels when they rate, but do make a holistic judgment (Magnan 1988).

However, the fact that a particular standard may be found to reflect decisions that can be criticized—the fact that a standard is "arbitrarily" set (Lantolf and Frawley 1985)—is not in itself an argument against it, since all standards are arbitrary value judgments, whether they are fire standards, health standards, or environmental standards (Popham 1978; Hambleton 1978; Cronbach 1961, cited in Davies 1988; Linacre 1992). Indeed, the arguments voiced against setting common standards, when there is a clear need and a wide consensus that they should be developed, remind me of the inability of policymakers to agree to stop global warming because researchers have not finished explaining it yet.

As Popham puts it,

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

Developments since 1978 do enable the "arbitrariness" or judge-subjectivity to be reduced, but not removed.

A scale of proficiency has two axes: horizontal (categories, which might be seen as a validity issue) and vertical (levels or bands, which might be seen as a reliability issue). In other words, there is a description issue (that the categories employed are related to a model of competence), and there is a measurement issue (that since everyone will treat the scale as if it were linear, it should be related to a model of measurement). In considering these two aspects, there is an equally important third aspect, a feasibility issue. Is the model a practical one? The fact that a model has sound academic foundations—that it reflects what applied linguists currently think they know or do not know about the nature of language proficiency—does not necessarily mean that it will work. An "arbitrary" judgment has to be made.

The essential problem is that even in the physical sciences there are no absolutely fixed categories or fixed scale values (Linacre 1989). All data is related somewhere to a theory, to a model that tries to take account of information given by existing data; and all data is collected, structured, and simplified by persons, and so affected by the theory or mental frame of reference they have. The problem with regard to language and mental measurement is, of course, far greater than for the physical sciences, but it is a difference of degree and not of kind.

The following two sections will consider the description problem and the measurement problem. The first section concludes that while one can take account of state-of-the-art models that try to describe communicative language proficiency/competence, it is in fact very difficult to establish the construct validity of a model of communicative competence through classic quantitative-analysis methods like correlations, factor analysis, and multitrait-multimethod analysis. The second section, on measurement issues, discusses the kinds of problems involved with the subjective judgments that are an inevitable part of scale or framework development, and proposes that a state-of-the-art measurement model—the many-faceted Rasch model (Linacre 1989)—can provide a pragmatic solution. Indeed, recent studies of the ACTFL Guidelines, perhaps spurred by the objections about "arbitrariness," have used this methodology to demonstrate that if the hierarchy of the ACTFL scale is arbitrary, this arbitrary judgment is shared by a far wider and more diverse range of people than critics may think—including naive native speakers (Dandonoli and Henning 1990; Kenyon and Stansfield 1992).

The Linacre many-faceted Rasch model can yield rich information about (1) how different groups of partners in the overall system—teachers, students, potential employers—interpret information and interact with ways of describing competence; and (2) how different components of competence interact in performance on tasks at different levels. Such information can inform decisions taken in the process of developing a scale about how to describe different aspects of competence, at different levels, for different users (the description issue), as well as addressing the measurement issue. Such an approach cannot address what is taken traditionally to be the key issue in construct validity, namely convergent and divergent validity—that different ways of testing/describing the same thing should get very similar results, while similar ways of testing/describing different things should get very different results—but as is described in the next section, that approach does not seem to be getting very far. When it is applied to underlying aspects of competence, the results are inconclusive (see discussion below); when it is applied to the four skills (Dandonoli and Henning 1990), it does not tell us a lot more than that the four skills were found to be separate, or that test method effects got in the way of finding that they are separate. The Rasch model can provide information about how subdomains within a skill (e.g., writing different kinds of text) differ, which could be the basis for a scale based upon "language activities" rather than the four skills (North 1992b; North et al. 1992); indeed, these distinctions have been found to be more significant in explaining student test performance (Pollitt and Hutchinson 1987, discussed below).

An approach exploiting a many-faceted version of the Rasch model, while not empirically demonstrating strict construct validity, could take account of Messick's broader definition of construct validity to include relevance/utility, value implications, and social consequences (Messick 1989), by identifying what factors appear relevant and measuring how different partners in the evaluation network put value on them in order to arrive at a system that has its "arbitrary" decisions resting on a theoretical and empirical basis.

2. THE DESCRIPTION ISSUE

Models of Communicative Language Competence/Proficiency

Stern lists a number of interpretations of language proficiency (Stern 1983):

— single-concept approaches—not held seriously since Oller's retraction or rewording of the unitary competence hypothesis (Oller 1976, 1983);

— binary concepts like Cummins's BICS/CALP (basic interpersonal communicative skills/cognitive academic language proficiency)—decontextualized language trapped by school tests (Cummins 1979, 1980, 1983);

— Canale and Swain's classic model (Canale and Swain 1980, 1981; Canale 1983); and

— multiple categories such as those put forward in a Council of Europe context by Van Ek and Trim and by Carroll, to whom one should add Morrow (Van Ek 1975, 1986; Van Ek and Trim 1990; Carroll 1978, 1980; Morrow 1977).

Canale and Swain's ideas, Cummins's BICS/CALP, and Bialystok's distinction between explicit and implicit learning (Bialystok 1982, 1986) were used as the basis for the Development of Bilingual Proficiency Project (Harley et al. 1987, 1990), which will be briefly discussed below.

Van Ek and Trim's ideas have since developed in a direction very similar to those of Canale and Swain (Van Ek 1986) and have been incorporated into a second edition of *The Threshold Level* (Van Ek and Trim 1990).

Carroll's ideas have been operationalized in the British Council's ELTS—now revised and called IELTS, the International English Language Testing Service, and administered jointly by the British Council, the University of Cambridge, the International Development Program of Australian Universities and Colleges (IDP), and Australian Education Centi es (Ingram 1990; Ingram and Clapham 1988; Westaway, Alderson, and Clapham 1990)—and in the English Speaking Union Framework Project, which calibrated British EFL exams to a common nine-band yardstick of descriptors (Carroll and West 1989).

Morrow's blueprint of *Techniques of Evaluation for a Notional Syllabus* for the

Royal Society of Arts (Morrow 1977) developed into RSA examinations in the Communicative Use of English as a Foreign Language (CUEFL)—now called the Certificate of Communicative Skills in English and run by Cambridge (University of Cambridge/RSA 1990)—and the range of foreign language examinations offered by the RSA Examinations Board (RSA 1989), developed from the CUEFL and the experience gained during the graded-objectives movement (Page and Hewett 1987).

There is a considerable amount of overlap among what might be considered the three leading theoretical models—those of Canale and Swain (Canale and Swain 1980, 1981), of Van Ek (Van Ek 1986; Van Ek and Trim 1990), and of Bachman (Bachman and Palmer 1982; Bachman 1987/88, 1990b). However, the distinction among pragmatic, discourse, and sociolinguistic competence is not always clear (Schachter 1990). At least some people have difficulty keeping apart all the "socio" categories—of which Van Ek has four in his original 1986 version (Council of Europe 1991, p. 59)—and all three groups of authors have shuffled the grouping of categories in succeeding versions of their models.

Bachman and Palmer's original multitrait-multimethod study of grammatical, pragmatic, and sociolinguistic competence (the traits) through a modified ACTFL interview, a writing sample, a multiple-choice test, and a self-rating (the methods), using confirmatory factor analysis, found a higher-order general factor plus two trait factors that they called grammatical and pragmatic competence (Bachman and Palmer 1982). The multitrait-multimethod analysis they used (Campbell and Fiske 1959) has been used by psychologists as an empirical test of construct validity: a test that the thing supposedly being tested—the trait or construct—exists and is trapped by the tests. In the method, convergent and divergent validity are established if the different measurements of the same trait (say, grammatical competence) correlate more highly than the same methods (say, multiple-choice tests) across the different traits.

The Development of Bilingual Proficiency (DBP) Project at Toronto tried to take things a step further by again using confirmatory factor analysis in an attempt to validate the Canale and Swain model empirically (Canale 1983 version): grammatical, discourse, and sociolinguistic competence (presumably, like Van Ek, seeing strategic competence as "compensatory"), plus Cummins's BICS/CALP and Bialystok's implicit learning/explicit learning distinction. The results were disappointing, failing to support the hypotheses.

Bachman attributes the failure to the construction of the test, suggesting that the instruments were more complex than the traits they were trying to measure; the model got mixed up in its operationalization. Second, he criticizes the use of a rotation approach—orthogonal rotation, normally used when traits are not expected to correlate. Third, he suggests that the interference of test method probably accounted for as much variance as the traits being measured, even if they had not got mixed up anyway (Bachman 1990a).

Schachter, at the same symposium as Bachman, argues on theoretical grounds for a basic grammatical/pragmatic distinction, and argues again that the model was not conceptually clear and was therefore imperfectly operationalized—hence the finding of one large factor. Attributing the large factor to a general proficiency, as Oller had done (Oller 1976), was to miss the point: the problem was in the conceptualization (Schachter 1990).

Paulston is a little more blunt than either Bachman or Schachter:

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

Bachman's conclusion is that, in terms of the history of empirical research into the nature of language proficiency as outlined by the DBP investigators (Allen et al. 1983, pp. 55–58), just as Oller's research on a global "g" factor and his retraction (Oller 1976, 1983) brought to an end the era of exploratory factor analysis, so the DBP study brings to an end a second period characterized by "increasingly complex and comprehensive frameworks of language proficiency" and by "sophisticated (and exhausting) research designs and statistical analyses." He continues:

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

Factor analysis, especially using the multitrait-multimethod approach used by Bachman and Palmer (Campbell and Fiske 1959), has long been the established method in behavioral psychology to establish construct validity empirically. Yet other studies, concerned with the direct rating of proficiency, also suggest that attempts to trap underlying parameters of language ability through quantitative methods like factor analysis and multitrait-multimethod analysis have only limited chances of success.

For example, two studies using factor analysis of performance ratings from behavioral scales of work performance (Norman and Goldberg 1966; Kavanagh, MacKinney, and Wolins 1971) suggest that "a factor analysis of ratings tells us more about the cognitive structure of the raters than the behaviour patterns of the ratees" (Landy and Farr 1983, p. 155). In other words, the parameters people see may reflect more the way they think and less what they are looking at. This reinforces findings from two studies on "halo effect" (transfer of judgment from overall holistic rating to rating for specific categories, or between categories); these studies suggest that "observed attribute intercorrelations may be at least partially a product of raters' conceptual schemes as well as of the true intercorrelation between traits" (Cooper 1981, p. 223, summarizing Passini and Norman 1966 and Norman 1963).

Borman therefore argues that it is a mistake to use the ratings of different "partners" as the methods in a classic multitrait-multimethod analysis to establish the validity of performance ratings, expecting them to display "convergent validity" (i.e., to agree on ratings), since in work performance the partners (supervisors, peers, subordinates) have legitimate separate perspectives and may well be concerned with different aspects of performance. Why should their views converge (Borman 1974; Landy and Farr 1980)? Ratees in different roles may exhibit different "true scores"; and different measurement sources, particularly raters with different perspectives, "may capture different aspects of the total criterion construct space" (Lance, Teachout, and Donnelly 1992, p. 447).

Strong halo effects are shown in a recent language study (Hamp-Lyons and Henning 1991) attempting to use multitrait-multimethod analysis to validate definitions of rating qualities designed to give communicative writing profiles (with qualities as traits and raters as methods). The raters (all the same type of partner) could not keep the rating factors (competence parameters, qualities of performance) sufficiently separate; their judgments on the different qualities were related to one another and did not confirm the independent existence of the qualities. This repeats the finding of Yorozuya and Oller's factor analysis using rating scales without any definitions; they claimed strong halo effect whether one sample was rated on all the parameters before looking at the next sample, or whether all the samples were rated first for one parameter, then for the next, and so on (Yorozuya and Oller 1980). That research itself repeated results from a series of studies conducted between 1956 and 1968 in work evaluation (cited by Landy and Farr 1983, p. 149).

A significant study by Pollitt and Hutchinson describes a three-component writing-skills assessment approach focusing on *appropriacy* (equated with sociocultural competence), *ideas structuring and selection* (discourse competence), and *expression* (grammatical competence), with the sociolinguistic element being determined by the context, audience, and purpose of the tasks—a letter, a report, a newspaper article, a story, an opinion (Pollitt and Hutchinson 1987). Two forms of analysis were used—first a traditional correlational analysis, and then a Rasch analysis. The correlations showed that the ratings for the underlying competences (performance qualities) were strongly interrelated, but that performance on one task appeared to be almost completely independent of performance on another. The authors conclude:

> The results underline the importance of including a wide enough range of language functions in writing tasks in any comprehensive language assessment, while at the same time they suggest that the particular model of competence used *(to rate the performance on different subscales)* may not be too important. (Pollitt and Hutchinson 1987, p. 90)

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

To summarize: One can expect that multitrait-multimethod analysis will continue to be used as a prime methodology for establishing the construct validity of language tests—particularly to establish that a test traps the trait concerned (e.g., listening) more than the method (e.g., deducing the correct multiple-choice alternative), as in Bachman's series of studies. However, the technique appears to have limited application for the development of framework descriptors, for two reasons:

— It is very difficult to operationalize the theoretical constructs in tests that keep them separate; they appear not to be homogeneous traits. The Development of Bilingual Proficiency Project members conclude that the elements of the Canale and Swain model—and, one might infer, the Van Ek and Bachman models—are not unitary, pure, or homogeneous: "Each trait may be made up of many different components, and there is no reason why all the components in a trait must correlate" (Harley et al. 1990). This point is extended to rating scales in a recent evaluation of multitrait-multimethod analyses: "If items from the same scale actually reflect different traits, or items from different scales actually reflect the same trait, then scale scores cannot be interpreted in terms of trait and method effects" (Marsh and Hocevar 1988, p. 108, cited in Lance, Teachout, and Donnelly 1992, p. 439).

— It is very difficult for raters with the same perspective to keep such constructs separate when rating performance because of (1) the fact that the analysis may show more what was in the raters' heads than what was in the performances; (2) halo effect/holistic rating; (3) the fact that the traits may be nonhomogeneous and therefore genuinely interrelated—so-called true halo; and (4) the fact that the student performances on a scale of proficiency may very well yield predominantly flat profiles, since the steps on most scales are relatively large in relation to learning development.

What is more, Pollitt and Hutchinson's experiment suggests that the shape of the underlying model of competence is less significant than the range of tasks to be performed. Their study suggests that it might be more useful to take a sociolinguistic ordering by context of use (task, language activity) as the principal horizontal axis in a descriptive framework system, rather than the underlying competence parameters and/or the particular rating categories used. This is in fact the approach taken in the business-oriented IBM France and ELTDU grids of subscales (IBM 1978; ELTDU 1975) and proposed by North as a option for a common framework (North 1992b; North et al. 1992).

It may be just as well that the underlying model of competence used in rating qualities appears to be less significant than the range of sociolinguistic tasks or contexts. Although there is considerable agreement that underlying competences of the type discussed exist, there is no particular reason, considering the arguments above, why they should be directly observable and ratable any more than there is reason to believe, given the disappointing results of the Canadian Development of Bilingual Proficiency Project, that it is possible to write test items that trap them. This suggests to me that underlying traits like linguistic, discourse, sociolinguistic, and sociocultural may not be appropriate bases for assessment scales, as for example Bachman has proposed (Bachman and Savignon 1986; Bachman 1987/88, 1989, 1990b). Furthermore, it seems so difficult to define them that an attempt to do so (Bachman and Palmer 1982, 1983, cited in Bachman 1987/88, 1989, 1990b), as Brindley comments (Brindley 1991, p. 11), seems forced into precisely the kind of juggling of qualifiers like *some* and *many* that Alderson has noted (Alderson 1991a), and that lead to the kind of confusing, word-processed alternatives criticized by North (North 1992b).

Bachman comes to the conclusion that a common scale should be expressed in abstract terms in order to avoid defining it "with reference to the performance of different groups of test takers," since it is "virtually impossible to define criterion levels of language proficiency in terms of actual individuals or actual performance" because "zero" and "perfect" language competence do not really exist (Bachman 1989, pp. 254–56). This is a very strict interpretation of Glaser's original concept of criterion-referenced assessment (see point 8 in the list given in the first section). Psychometrically, the requirement of such an interpretation can in fact be met by applying a Rasch model during the development of a scale, since "zero" is to be found at the center of the scale while "perfection," infinity, is to be found at the two ends. Divorcing the development of such a scale from the description of actual performance seems in any case to run counter to Glaser's intentions. According to Glass, Glaser chose the term *criterion* because of its classical psychometric meaning as "a scale formed by the observation or recording of behaviour which the psychometric instrument is to predict" (Glass 1978, p. 242).

One alternative to an abstract approach based on underlying competences is what might be called a pragmatic approach based on operational models of competence, the kind of criteria for degrees of skill suggested by Carroll and Morrow and used in the suites of communicative examinations becoming available in Europe (e.g., University of Cambridge/RSA 1990). These approaches take account of theories of underlying competences but regroup them in order to focus on features that are more observable—or features that, it is felt, should be highlighted with regard to a particular task. It does not fully solve the problems of vagueness and word-processed symmetry mentioned above, but it does make things a little easier.

The approach to the assessment of group interaction developed for Eurocentres U.K. (North 1986, 1991, 1992a), for example, splits linguistic and sociolinguistic competence into *range* (after Carroll 1980) and *accuracy* (which includes appropriacy, as with ACTFL); it has Bachman's psychophysiological competence (Bachman 1990b) plus Faerch and Kasper's preplanning and processing (Faerch and Kasper 1983) under *delivery*; and it puts discourse "challenge" (getting help when you don't follow—Burton 1980), other strategic competence, "collaborative moves" (Barnes and Todd 1977), and sociocultural competence under *interaction*. (Range, Accuracy, Delivery, Interaction, plus Overall = RADIO.)

Rating categories seem in fact to vary just as much as discourse-analysis and interaction-analysis categories do. There are myriad factors, but one can work only with a few, and so people group them in different ways. Institutions develop their own criteria and train raters to use them, developing "schools" in the process. Some schools seem to think they have an exclusive definition of proficiency. Most, however, recognize that there are many routes to the same goal. In British EFL, teachers often switch between systems as they work for their school or examining boards, or they give their students a "mock exam."

Since rating categories are a metalanguage used to talk about competence, and since this is a very valuable experience for teacher development, it can be argued that the competence categories used—whether of the underlying variety like "sociocultural competence" or of the operational variety like "range"—should have relevance for the people who are expected to use them, and preferably should be developed empirically with them, rather than being standardized in a common framework. It is, of course, possible to generalize across systems and create "supracategories," as North and Page have recently done in a synthesis across thirty-five proficiency scales (North et al. 1992); but such a generalization is merely a lingua franca to people's metalanguages.

Behaviorally Based Scales in Work Performance

Developing assessment scales with the kind of people who are going to use them was the approach pioneered by Smith and Kendall, who appear to have developed the first defined or "transparent" assessment scales outside the U.S. Foreign Service Institute (Smith and Kendall 1963). They were reacting against a practice whereby abstract categories (traits) determined by psychologists on the basis of intuition or factor analysis were parachuted into hospitals to be used by head nurses in rating their juniors. Smith and Kendall's motivation was the unreliability of the resultant ratings, and they developed the first form of what are called, generically, behaviorally based rating scales. The particular form they invented is called either behavioral expectation scales or, more usually, behaviorally anchored rating scales, because the rating scale has "anchors" of expected performances—behavior you would expect to observe.

The history of the use of behaviorally based scales in work evaluation shows some parallels with the history of the development of scales of proficiency in language teaching, with ıegard to the ways people have tried to give meaning to numbers, to describe the features being assessed.

Before the arrival of nımerical rating scales in the aftermath of World War I, all one had was weighted marks for undefined characteristics or dimensions—as indeed one still sees in many foreign language examinations, such as the new Diplôme Elémentaire de Langue Française.

Two classic forms of numerical rating scale are reproduced in Figures 1 and 2. The first has a short definition of the "dimension," "trait," or "skill" concerned, plus an indication of the meaning of the two ends of the continuum; the second just labels the dimension. Such rating scales do not have definitions attached to the steps on the scale itself. Both types illustrated have been used in language examinations. For example, the FSI oral interview used such a simple scale in the 1970s for the performance of the factors accent, grammar, vocabulary, fluency, and comprehension (Wilds 1975, p. 38):

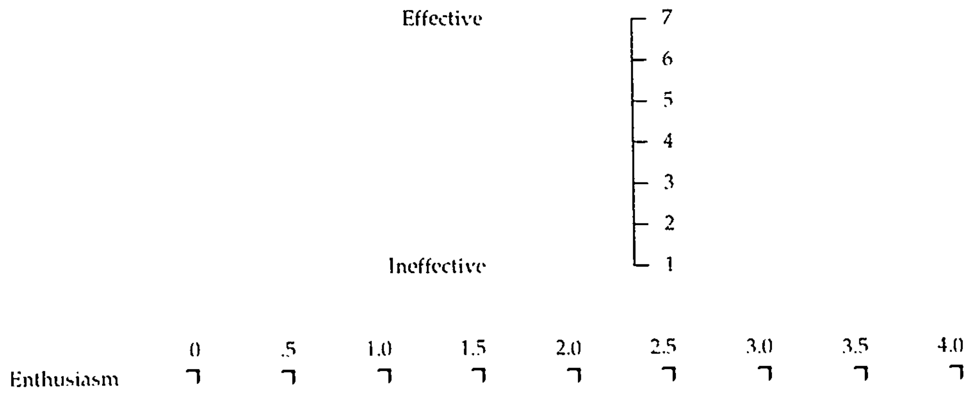
1. ACCENT foreign ___ : ___ ___ : ___ ___ : ___ native

The Goethe Institute's Kleines Deutsches Sprachdiplom still uses a numerical marking scale with just a label for each dimension.

Figure 1

Administrative Skills

Planning ahead; organizing time efficiently; completing paperwork accurately and on time; keeping track of appointments; not wasting time.



Source: W. C. Borman, "Behavior-based Rating Scales," in *Performance Assessment: Methods and Applications*, ed. R. Berk (Baltimore: Johns Hopkins University Press, 1986), p. 102.

The first attempt to provide detail about the kinds of behavior associated with different parts of the continuum represented by the scale was the development of the "graphic rating scale" (Paterson 1922; Freyd 1923). The original graphic rating scales were a continuous line between two points. The dimension being measured was described in a general definition at the top of the scale, and behavior associated with different parts of the continuum was described in short definitions called cues. The cues were spaced equidistantly along the continuum and connected to it so as to present scale steps. However, raters were not asked to pick the most appropriate cue (or scale step); rather, having decided which cues were most appropriate to the performance being rated, they were asked to mark a point on the continuous line itself, which would necessarily be between cues.

The next significant development was to arrange the scale vertically rather than horizontally, thus allowing more room for longer and therefore more precise cues (Champney 1941). The difficulties of formulating precise cues, and the danger of vague relative language that has recently been criticized in relation to scales of language proficiency (Brindley 1991; Alderson 1991a—see point 4 in the list given in the first section of this paper), were recognized in 1941:

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

Champney's second innovation was to promote a multidimensional model with a scale for each dimension; all subjects (talking of five or six) would be rated on the same dimension before going on to the next dimension.

The weak points of the graphic rating scale methodology, even incorporating Champney's innovations, were deciding the dimensions, selecting or designing the cues, and deciding what scale value to give the cues on each dimension. To assign values, Champney used a rank-ordering rater-agreement task still often used in scale construction.

Smith and Kendall therefore developed a rigorous methodology of cross-checking workshops and item analysis for identifying dimensions that made sense to the people who were to use the scale, for selecting the cues from a pool offered on the basis of consistent interpretation, and for giving scale values to those most appropriate cues. These behavioral cues "anchored" rating observations to the continuum—hence the name behaviorally anchored rating scale, or BARS (Smith and Kendall 1963).

BARS scales take various appearances, a classic simple one appearing in Figure 2. Note that there is no attempt to give a behavioral anchor to each step on the scale, nor do the anchors line up against scale points exactly; they are situated in the band between the points at the mean rating they received in the development workshops. The rater thinks of the behavior that has been observed in relation to the anchors, deciding whether this represents a higher or lower level on the continuum, and selects the most appropriate scale step for the rating.

A further innovation by Smith and Kendall was to expand the qualitative description of the dimension that appeared at the top of the scale (see Figure 2) into three qualitative, more abstract descriptions: for a very high performance, for a very low performance, and for an average performance—or minimum adequate competence—in the middle. These three longer paragraph descriptors were then put on the left of the scale. A variation on the original Smith and Kendall approach to qualitative descriptors is given by Landy and Farr (1983, pp.

Figure 2

Salesmanship Skills

Skillfully persuading prospects to join the navy; using navy benefits and opportunities effectively to sell the navy; closing skills; adapting selling techniques appropriately to different prospects; effectively overcoming objections to joining the navy.

9

A prospect stated he wanted the nuclear power program or he would not sign up. When he did not qualify, the recruiter did not give up; instead, he talked the young man into electronics by emphasizing the technical training he would receive.

8

The recruiter treats objections to joining the navy seriously; he works hard to counter the objections with relevant, positive arguments for a navy career.

7

When talking to a high school senior, the recruiter mentions names of other seniors from that school who have already enlisted.

6

When an applicant qualifies for only one program, the recruiter tries to convey to the applicant that it is a desirable program.

5

When a prospect is deciding on which service to enlist in, the recruiter tries to sell the navy by describing navy life at sea and adventures in port.

4

During an interview, the recruiter said to the applicant, "I'll try to get you the school you want, but frankly it probably won't be open for another three months, so why don't you take your second choice and leave now."

3

The recruiter insisted on showing more brochures and films even though the applicant told him he wanted to sign up right now.

2

When a prospect states an objection to being in the navy, the recruiter ends the conversation because he thinks the prospect must not be interested.

1

Source: W. C. Borman, "Behavior-based Rating Scales," in *Performance Assessment: Methods and Applications*, ed. R. Berk (Baltimore: Johns Hopkins University Press, 1986), p. 103.

62–65). In that variant the continuum is divided into three sections or broad ranges of level, with the qualitative summary statement covering the whole of each range rather than points at the two ends and middle; the behavioral anchors are also grouped in these three broad levels.

The BARS combination of, in our terminology, *task* information in the behavioral cues or anchors, plus *qualitative* information for broader levels, helped increase transparency. People could see what was being talked about. A rash of empirical studies conducted in the 1970s were inconclusive as to whether such transparency led to improved accuracy and consistency of the ratings themselves (for reviews, see Jacobs, Kafry, and Zedeck 1980; Kingstrom and Bass 1981; Borman 1986). Nevertheless, the potential of a metalanguage for rater training, ratee feedback, personnel training, job analysis, policymaking, and so on made BARS an increasingly popular evaluation format.

A major consideration in the BARS approach is that the behavioral anchors refer to very concrete, specific examples of behavior—the kind of thing a person at this level could be expected to do. Indeed, the original name of the Smith and Kendall invention was behavioral expectation scale (BES). The equivalence of such an approach for a scale of language proficiency would be a numerical scale with the anchors being tasks the student could be expected to perform at each of the levels on each of the dimensions. A language certificate statement like:

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

might form the basis of an anchor like:

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

The appeal of BARS is the concreteness of the anchors. The main difficulty with the approach is that the raters have to judge where the behavior observed "fits" on the scale. To do that, they have to infer how a person would behave in a specific situation, and some people experience difficulty doing that. Notice that the BARS approach of focusing on one very specific behavior to anchor a level of performance would mean that the more general part of the certificate statement for this level ("can write on a range of subjects") and another specific behavior mentioned ("can compose straightforward formal letters so that . . .") would both be omitted from the scale.

This demonstrates the main problem people had with BARS: the examples tend to be too specific, and it can be difficult to generalize from them. Although BARS scales continue to be used extensively, the response to these difficulties has been the development of two rival successors, which parallel very closely the directions taken in applying the ideas behind behavioral scales to language learning: behavioral observation scales, and behavior summary scales.

Behavioral observation scales (BOS). In the BOS approach, a long list of tasks in the domain is provided, with each task separately rated on a numerical scale, usually 0 to 5. One variant would be just checking off the tasks (yes/no). Another would be using what is called a modified standard scale—zero for an average performance, minus for a low one, and plus for a high one.

This is the kind of quantitative approach used in the graded-objectives movement. It is closely related to behavioral objectives in nonlanguage vocational education (typing, machine skills), the so-called competence-based approach, the "mastery learning" interpretation of criterion-referenced assessment. The trouble is that it does not place the student on a continuum in an overall framework; it is difficult to generalize about competence on the basis of checked-off assessments of performance on a list of tasks, particularly when, as is often the case, there is no comment about the quality of the performance. Such an approach could, however, be used within a defined broader level as a form of continuous assessment, and in most language applications this seems to be implicitly the case, since the applications tend to be in programs aiming to get students up to the threshold level.

Behavior summary scales (BSS). In the BSS approach, anchor paragraphs are written that are representative of and common to the broader range of behaviors, incidents, and subskills that are scaled at each level, and more abstract comment is included. In other words, the kind of abstract comment about the quality of performance that started appearing on some forms of BARS is extended with examples of specific behaviors that are intended to be representative. The scale will probably have three or four subscales on different performance aspects, and it may group narrower numbered levels into broader defined levels (Landy and Farr 1983, pp. 104–9). BSS scales thus take criterion-referenced assessment to mean not declaring someone's mastery of specific points in a domain, but rather identifying someone's stage of development on a continuum (Hambleton 1988; Berk 1988; Glaser 1963).

One way the BSS approach has developed in the language field is the LSP (language for specific purposes) scales of ELTDU and IBM France mentioned at the beginning, with a rating for each level for a set of specific contexts of use (language activities); ELTDU has twenty-six such subscales, and neither ELTDU nor IBM France has a global scale. Another application is the scaling of different aspects of performance in assessment subscales for the degrees of skill required (Carroll 1980; Carroll and Hall 1985; Carroll and West 1989; IELTS; University of Cambridge/RSA 1990; RSA 1989; Eurocentres scales of language proficiency, assessment versions).

The BOS and BSS approaches differ in presentation, but all three types (BARS, BOS, and BSS) tend to share the same development technique—usually simplified and less rigorous derivations of that employed by Smith and Kendall—and as a result they tend to produce similar results. During the 1970s there was a whole series of inconclusive format-comparison studies, some showing the one format to be superior, some the other. Apart from methodological problems that make comparison difficult, if not impossible (Kingstrom and Bass 1981), the inconclusiveness would appear to be due to the fact that the difference is mainly a presentational one, and superiority/inferiority is probably due more to rigor/lack of rigor during development (Landy and Farr 1983; Borman 1986). Landy and Farr as well as Borman have therefore called for a moratorium on format-comparison research and a concentration on deciding at what level to anchor the anchors—which through inadequate item analysis often systematize the random error they are designed to exclude. The anchors should be anchored through a method based in psychometric theory, and a lot of the problems experienced with behaviorally based scales can be traced to the fact that they are not (Landy and Farr 1983).

Where the approaches are felt to differ is with regard to giving feedback. BSS scales, with the three or four subcategories, provide a metalanguage for feedback and a justification for decisions in a way that the seeming arbitrariness of the lists in BOS or the very specific selected anchors in BARS do not.

Historical Development of the ILR/ACTFL Guidelines

The FSI/ILR/ACTFL system has developed over the same period of time as behaviorally based scales for work performance, and it has probably influenced, directly or indirectly, virtually all other systems except ELTDU—which claimed inspiration from the Council of Europe threshold-level approach (ELTDU 1975). It is therefore interesting to see how the two approaches, FSI and behaviorally based work evaluation, developed in parallel. What is fascinating is that the FSI seems to have jumped straight from graphic rating scales to behavior summary scales for its reporting instrument some twenty years before they came into use in other disciplines, but for the rating itself the FSI seems to have stayed with graphic or Likert scales—categories with little headings like those on all opinion polls (Myford 1991):

Very	Moderately	Slightly	Slightly	Moderately	Very
Well	Well	Well	Poorly	Poorly	Poorly

Outside the FSI/ILR group, many U.S. studies in language testing (e.g., Yorozuya and Oller 1980; Davidson and Henning 1985) continued to use simple nondefined scales, as many language examinations still do. In the United Kingdom, on the other hand, the BSS approach has developed into a large number of systems, and the BOS approach has informed the graded-objectives movement and competence-based approaches like that of the Scottish Vocational Educational Council.

Outside the language field, rating in work performance seems to have been done on purely graphic or numerical scales—sometimes with Likert labels for the steps—until Smith and Kendall's breakthrough in 1963; and since then, while work performance evaluation has tended to adopt behaviorally based scales, the majority of rating scales discussed in the literature on "rating scale analysis" still seem to use nondefined graphic, numerical, or Likert scales. The example of a

Likert scale given above is from a very recent study using a Rasch analysis of ratings that will be referred to later.

In the early 1950s the FSI team took as its starting point a graphic scale and a definition of a "useful" level of competence arrived at in 1952, and then devised short paragraph-length definitions for each level, which seem to have been used right from the start for the reporting scale (Liskin-Gasparro 1984).

According to Lowe, the linguistic-theoretical base of the rating procedure was Bloomfieldian structuralism, and the psychometric base was criterion-referenced theory, which demanded zero and perfection as the two polar points in order to utilize Osgood's "semantic differential" to state the amount present of one of a pair of polar terms—hence the notorious "educated native speaker" as Level 5 (Lowe 1985). During the interview, the rater considered five factors (accent, grammar, vocabulary, fluency, and comprehension) and marked grades on a continuum (see diagram on p. 18, above). Accent, for example, could be rated as being relatively foreign-sounding or relatively native-sounding (Lowe 1985). Apparently, brief verbal descriptions of the steps quickly led to a Likert scale that could be made available as a grid defining the factors for each level (Lowe 1985; Clark and Clifford 1988). The grades for the factors, however, did not automatically add up to the global grade; the rater made a holistic judgment, informed by the factors. Indeed, the factors have been shown to operate with different weightings at different levels (Clifford 1980). Incidentally, these three characteristics—working with a nondefined checklist, having definitions available if necessary and for training, and making a holistic judgment rather than doing arithmetic—are all paralleled in the Eurocentres approach to testing group interaction, with the additional characteristic of distinguishing between an initial holistic impression and a confirmed final judgment made by comparing the analysis by factors with the original impression (North 1991, 1992a).

The original FSI reporting definitions were paragraph-length statements:

> *Level 1 Elementary Proficiency.* Able to satisfy routine travel needs and minimum courtesy requirements
>
> Can ask and answer questions on topics very familiar to him; within the scope of his very limited language experience can understand simple questions and statements, allowing for slow speech, repetition or paraphrase; speaking vocabulary inadequate to express anything but the most elementary needs; errors in pronunciation and grammar are frequent, but can be understood by a native speaker used to dealing with foreigners attempting to speak his language; while topics which are "very familiar" and elementary needs vary considerably from individual to individual, any person at Level 1 should be able to order a simple meal, ask for shelter or lodging, ask and answer simple directions, make purchases and tell time. (Wilds 1975; Clark and Clifford 1988; 112 words, excluding title)

The definition for each succeeding level was shorter and a little more abstract than the previous one; only Level 1 went in for specific behavioral expectations ("any person at Level 1 should be able to . . .") reminiscent of the BARS approach discussed above; higher levels took a more qualitative BSS approach.

The major revision of the FSI descriptors to produce the ILR descriptors took place at the same time that the first "provisional" ACTFL Guidelines were developed, with a slightly revised version being published in 1991 (ILR 1983; DLI 1991). The ILR descriptors have definitions for all four skills, whereas the FSI definitions had only speaking and reading, and the "plus levels" between the original 0 to 5 are also fully defined. The speaking definitions are now split into two: a general part, and then a longer "examples" part that describes what typical Level 1s can do. Assessor-oriented information about the quality of the performance the Level 1 might give in the interview is totally mixed in both parts of the definition with information about the tasks a Level 1 could be expected to handle. The descriptions still tend to get shorter for each succeeding level. The definition for Level 1 has a distinctly more negative feel than the FSI descriptor, and in the 1983 version it is 291 words long (1991 is very similar).

The 1982 provisional ACTFL Guidelines have a layer of generic descriptors applicable to all languages, and then language-specific descriptors. They split Level 1 into intermediate low and intermediate mid (Level 1+ becomes intermediate high). The generic descriptors are more in the style of the FSI descriptors than the ILR ones. Intermediate low is 126 words long, and mid is only 45 words long, again excluding the title, which is actually now the first sentence. There is very little about quality of performance. The language-specific descriptors repeat the generic descriptor and exemplify it with the type of errors intermediate low students make. Both intermediate low and mid are about 180 words long. In the 1986 revisions, the generic descriptors for different levels have a more standardized length of about 80 words.

Compared with assessor-oriented scales developed in Britain by Carroll and others (Carroll 1980; Carroll and Hall 1985; Carroll and West 1989), by the IELTS revision team (Westaway 1988; Westaway, Alderson, and Clapham 1990), and by breakaways from it (Hamp-Lyons and Henning 1991), and compared with the Royal Society of Arts examinations, Cambridge examinations, and Eurocentres assessor-oriented scales, what is noticeable about the ACTFL system is the poverty of the assessment criteria—the descriptions of the quality of the performance. The ACTFL assessment grid (ACTFL 1989) has five very short keyword or note entries for each level, but three of these—global tasks/function, context, and content—have to do with knowing what general level is being talked about (user-oriented) and selecting which task to use (constructor-oriented) rather than with analyzing the language actually elicited (assessor-oriented). For analyzing the degree of skill in the language elicited, the factors are just "accuracy" and "text type," which for intermediate (no subcategorization) are defined as follows:

> *Accuracy:* Can be understood, with some repetition, by speakers accustomed to non-native speakers.
>
> *Text type:* Discrete sentences and strings of sentences. (ACTFL 1989)

This is not a very rich model; it seems to say, in effect, "If I can understand the person even though he or she just strings sentences together, then he or she is intermediate."

In an educational system that invests an average of approximately 260 hours per student in foreign language study, including college (Lambert 1992), such a description would presumably fit virtually everybody who is not a total beginner; Barnwell has made a similar point (Barnwell 1991). That might not matter, since there are subdivisions inside the level to show progress—though there is some suggestion in recent validation studies that the Guidelines/Oral Proficiency Interview operates as a threshold scale ("I can understand you; you are over the intermediate threshold") rather than as a serial scale distinguishing among all levels (Dandonoli and Henning 1990; Kenyon and Stansfield 1992).

Pollitt has criticized scales like the Australian Second Language Proficiency Ratings, the ACTFL Guidelines, and the English National Curriculum attainment targets for failing to give a definition of what constitutes an acceptable performance beyond merely describing a set of hypothetical tasks:

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

In reply it could perhaps be argued that there are indications that teachers/raters do not use detailed assessment criteria consistently—transferring judgment between categories in a halo effect (Hamp-Lyons and Hen~ing 1991; Yorozuya and Oller 1980); and in any case experience shows that if raters are not required to report final grades in terms of the categories but only to give a final holistic judgment, and if they use the system regularly, they "internalize" the criteria—that is, they stop looking at them (Jones 1985)—and therefore detailed criteria are a waste of time.

Certainly studies from the field of work performance evaluation give only very limited support to the idea that defined criteria increase the reliability of ratings. A number of studies show the presence or absence of defined labels making no significant difference to means or reliabilities (e.g., Finn 1972; McKelvie 1978), especially when the content of the rating tasks is familiar and individuals have developed a common perspective and a set of similar "preconceived and rather uniform judgment standards" (Finn 1972, p. 264).

On the other hand, a number of studies do show an improvement with definitions. Keaveny and McGann found that adding behavioral descriptors reduced halo effect and improved discriminant validity in a multitrait-multimethod analysis; raters found it easier to keep separate the dimensions being rated if they were given definitions (Keaveny and McGann 1975). Borman and Dunnette conclude that scales with defined behavioral descriptors clearly increase interrater reliability while reducing classic rater errors like halo effect and leniency. However, they calculate that the addition of defined descriptors increases reliability by only 5 percent of the variance at most, and they suggest that the true argument in favor of defined descriptors is rather the wealth of information they furnish about overall requirements and about individual performance in relation to those requirements (Borman and Dunnette 1975).

A comprehensive review of these comparative studies came to the conclusion that the addition of behavioral descriptors "can no longer be considered the means by which rating errors are minimized." From a quantitative, psychometric point of view, scales with descriptors are no better or worse than other methods, but the real potential is in the qualitative improvements that come from defining and giving feedback in relation to common goals (Jacobs, Kafry, and Zedeck 1980, p. 630). In a clarification of the original purpose of BARS, Bernadin and Smith declared that the aim had been to encourage accurate observation and recording of behavior in continuous assessment in order "to enhance future observation and to foster a common frame of reference in observers' ratings" (Bernadin and Smith 1981, p. 458).

With regard to language learning, even if halo effects with defined criteria can be high (Pollitt and Hutchinson 1987; Hamp-Lyons and Henning 1991), and even if the contribution of defined criteria to the reliability of the ratings is very modest, there are other arguments—from a validity standpoint—for including detailed criteria:

— *Feedback.* Detailed assessment categories supply a metalanguage for giving feedback on performance and suggesting areas to concentrate on. They can also be used in sensitization activities about the nature of language, and hence the nature of the activity of language learning. They can reveal that performance changes qualitatively with increasing proficiency, that assessment categories operate differently at different levels (Clifford 1980; Pollitt and Hutchinson 1987). They promote transparency.

— *Washback.* Unfortunately, virtually all support for washback effects seems to be anecdotal (Alderson 1991b). Nevertheless, the principle of what is defined as washback validity (Morrow 1986) or systemic validity (Fredericksen and Collins 1989) is very attractive.

At the very least, detailed assessment criteria sensitize teachers to the components of communicative competence and to the necessity for learners to acquire skills of language use by using the language in structured group interaction and pair work. Since teachers are good critics, such a training process also tends to lead to criticism of activities—materials evaluation in relation to the features of the language the material is supposed to elicit. Since the language generated by material depends on how it is used, how the activity is set up, the atmosphere—test method effect factors (Bachman 1990b, ch. 5)—this develops into discussion about the management of a communicative classroom and the teacher's role as observer. This, in turn, invites discussion of the phenomenon of teacher talk, or TTT (teacher talking time): Do the students ever get to acquire discourse skills through regular group interaction? Or is the classroom, like the interview, exclusively a place where questions are asked, answers displayed as knowledge, and responses patronizingly rewarded (Sinclair and Coulthard 1975, 1982)?

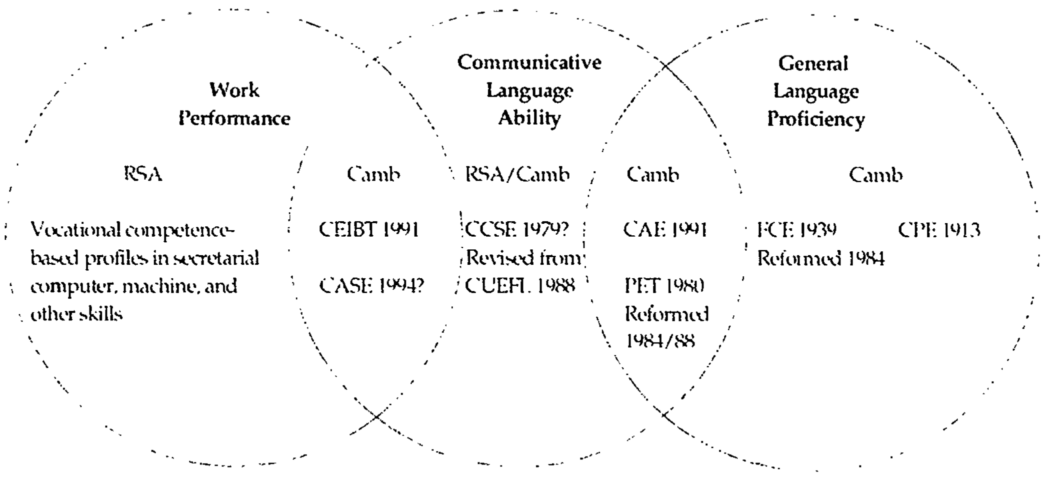
Assessment criteria detailing degrees of skill for different levels in different components of competence promote coherence among a view of teaching language as action, classroom organization and management, materials evaluation, and teacher training. Including criteria on linguistic factors (e.g., range, accuracy) as well as pragmatic or communication factors (e.g., delivery, interaction) encourages teachers—and therefore maybe learners—to see the necessity for structured practice of language *as* communication (actual use) as well as language *for* communication (study of usage).

The ACTFL Guidelines/Oral Proficiency Interview are said to have already had a substantial washback effect. Presumably this is because of the task information in the descriptors, translated into "situation cards" and descriptions of potential role-play scenarios at each level—what sort of things learners should be able to do, what sort of things they should be given a chance to do in the classroom.

But since the origin of this activity is a one-to-one interaction, the effect—as with very many examinations—can be to focus attention on sentence-level function responses (phrase-book language) and on situational dialogues rather than interaction, on communicative drills rather than communication. Both are useful methods that have been employed successfully for centuries, but in a supposedly communicative era there is something else as well, something that comes only from extended structured group interaction and has a focus on what has been called collaborative moves (Barnes and Todd 1977), discourse functions (Widdowson 1978; Van Ek and Trim 1990), and "ways in which turns are taken and given, addressees selected, floor held and interruptions accomplished" (Sinclair 1979). These features have been called "interactional competence" (Kramsch 1986), or communicative use, as opposed to usage (Widdowson 1978, 1979), of the language. The fact that the learner needs this kind of procedural experience as well as linguistic and sociolinguistic knowledge and experience, and the fact that assessment activities—and, by implication, scales of language proficiency—should take account of this in their design, is perhaps what is intended by the suggestion that the term *communicative language proficiency* should subsume the two terms *proficiency* and *communicative competence* (Bachman and Savignon 1986) in an assessment context.

Recently, certain of the U.S. government agencies that are members of the Interagency Language Roundtable have started to question whether the Oral Proficiency Interview system—"proficiency"—is delivering what they want. (Child, Clifford, and Lowe 1991; Walton 1992). The discussion thus far appears to concern defining the ends of the continua involved, with general proficiency at one end and job-specific performance testing at the other. In the paper by Child, Clifford, and Lowe, the relationship between proficiency and (work) performance is represented by circles, sometimes one inside the other, sometimes overlapping. Figure 3 is a development from the overlapping-circles diagram. It

Figure 3


Work Performance
Communicative Language Ability
General Language Proficiency
RSA
Camb
RSA/Camb
Camb
Camb
Vocational competence-based profiles in secretarial computer, machine, and other skills
CEIBT 1991
CASE 1994?
CCSE 1979? Revised from CUEFL 1988
CAE 1991
PET 1980 Reformed 1984/88
FCE 1939 Reformed 1984
CPE 1913

uses the interaction in recent years between the examinations of the University of Cambridge and the Royal Society of Arts to illustrate that it may be more helpful to think of three rather than two overlapping circles, and that "communicative language proficiency/ability" is this crucial middle circle. Some of the dates given are approximate.

Starting from the right-hand side of the diagram, the first EFL examination from Cambridge was the Certificate of Proficiency in English (CPE), which was apparently originally used as a barrier test for students who wished to follow a humanities course at the university. It seems to be a pattern in Europe that the first "serious" examination in a language should be developed with this goal. The Goethe Institute has one, the Kleines Deutsches Sprachdiplom (and as the name suggests, like Cambridge the institute has a higher one too); the Spanish have long had an exam at the same level as the Cambridge CPE and have recently developed initial and "basic" diplomas too; the French have developed the Diplôme Approfondi de Langue Française (DALF) as a university barrier test, with the Diplôme Elémentaire de Langue Française (DELF) as a series of staging posts at a lower level; and the Italians have just developed a CPE-level exam too. With the exception of DELF/DALF, all of these examinations take a traditional academic view of proficiency as manipulation of language usage, essay writing, plus culture. The exact mix differs from country to country.

The First Certificate in English (FCE), being developed later than the CPE and intended for a general public, is less academic. It was originally called the Lower Certificate but was conceived as a certificate at the first level of competence of public interest—the first level that might interest an employer. It was reformed in 1984 and moved slightly toward the center of the diagram—that is, it became slightly more communicative.

On the extreme left of the diagram are the Royal Society of Arts professional and vocational examinations, which have pioneered a competence-based approach, listing tasks the student can perform and word-processing this into a profile statement. The Scottish Vocational Educational Council system (SCOTVEC) applies exactly this BOS graded-objectives approach to its language certificates. At one point it used to be considered that LSP (language for specific purposes) developed once a learner had a general language proficiency, perhaps up to threshold level. Recently, however, it has been increasingly recognized that occupational users of a language may function with a very low level of achievement in discrete, specific contexts using a set of prefabricated chunks—islands of certainty around which they can improvise slightly. Organizing low-level vocational language objectives as a list of discrete behavioral objectives, as in SCOTVEC, is a way of recognizing the possibly discrete nature of low-level vocational language skills.

The Royal Society of Arts, however, did not assume that the behavioral-objectives approach it adopted in vocational studies would necessarily be appropriate to assess the whole continuum represented by language performance in EFL. The RSA therefore commissioned Keith Morrow of Reading University, the British cradle of the communicative approach, to write a specification for techniques for evaluating a notional syllabus (Morrow 1977). The Morrow report led to the development of the RSA communicative examinations for EFL. These have had several names but for most of the 1980s were known as CUEFL, examinations in the Communicative Use of English as a Foreign Language. Since 1988 they have been administered by the Cambridge Syndicate and are now called CCSE, Certificates in Communicative Skills in English.

The Morrow approach specified authentic tasks and gave detailed definitions of quality, the degree of skill expected at each level for a range of performance factors (BSS scales). These RSA examinations—though not taken by that many students, since Cambridge examinations were felt to have more currency—had a truly revolutionary effect on both communicative teaching and testing. They provided a wealth of task-based authentic material, with test exercises that looked like ideal teaching material, and they broke with the interview format by introducing two examiners, one of whom (echoing the original FSI format) operated as interlocutor and was usually known to the student, and the other of whom operated as the rater. Since the oral exam involved student-student interaction as well as conversation with the interlocutor, it was also task-based.

This format—communicative testing, general language performance testing—tests skills in using the language communicatively, skills that are generalizable to other, more specific contexts, rather than testing knowledge of the usage of the language (which, though of course in theory generalizable, may not actually be operationalizable). The RSA exams were developed as a reaction against the traditional proficiency approach adopted by Cambridge. The RSA diplomas are in fact preferred by all the large Swiss banks, because they certify skills that are applicable in different work contexts—but even the students from Swiss banks do not realize that and continue to prefer to take the First Certificate!

The Royal Society of Arts EFL exams, which have changed their name several times and are now in fact run by Cambridge, impacted right (into proficiency assessment) and left (into work assessment) on the diagram shown in Figure 3. The newer Cambridge exams in the traditional suite—the Preliminary English Test (PET) and the new Certificate in Advanced English (CAE)—have kept to the Cambridge house style and its belief that usage should also be tested directly because of its generalizability; but in the process they have adopted "communicative" tasks and "authentic" rather than literary texts, and more innovative interview formats. The new Cambridge exams aimed at people in the world of work—the Certificate in English for International Business and Trade (CEIBT) and the Cambridge Assessment of Spoken English (CASE)—both show the RSA influence. CASE, a diagnostic-profiling one-to-one oral examination that involves a variety of groupings of non-native and native speakers (Milanovic et al. 1992), is a particularly good example of how discourse concerns, procedural competence, and potential for washback and feedback are being taken increasingly seriously in assessment.

A criticism of the Royal Society of Arts EFL exams from language testers during the 1980s was that their claims to validity—certainly before they came into the Cambridge group—had an educational rather than a psychometric base. Construct validity was interpreted from theory, rather than demonstrated statistically, and reliability was not the top priority. This has also been the approach of the Council of Europe and Carroll's original ELTS specifications and test (Carroll 1978), basically saying, "Let's keep the cart before the horse"; support for this position is offered by the equivocal success of attempts to establish quantitatively the construct validity of models of competence outlined in this section, as well as by evidence of the absurdities that ignoring washback validity has led to in language testing (Savignon 1992).

However, in connection with the development of a system of descriptors for a common framework, if one wishes to develop a descriptive framework capable of expressing the outcomes from "hard data" collection in official examinations (i.e., if there is to be credibility to the "passport" in a possible European Language Portfolio related to the framework—Schärer 1992; Council of Europe 1992), it would be an advantage if the descriptive framework itself had an empirical psychometric base to support it (i.e., if it was rooted in a measurement theory). Otherwise, one will just have the "soft data," the "map" (Schärer 1992), to find one's way around.

The original ELTS was much criticized for the way in which the descriptors were arrived at, for failing to demonstrate psychometric properties (e.g., Skehan 1984), and for sometimes producing strange results. These criticisms led to an evaluation project (Criper and Davies 1988), a revision project (Alderson 1991a; Westaway, Alderson, and Clapham 1990), and a rebirth as IELTS (Ingram and Clapham 1988).

The difficulty of undertaking any empirical validation of the model of competence adopted—the description issue—has been discussed at some length above. It seems unlikely that anyone will statistically substantiate a model of competence in our lifetimes, even if workers in the field answer Bachman's call for a coordinated research program in the terms of his model. This is not necessarily such a problem in the context of the development of a common framework, since the consensus on components of competence is considerable and the strength of the Council of Europe project team is on the descriptive side. However, recent developments in measurement theory suggest that it is no longer necessarily true that you have to choose between reliability on the one hand and validity in its wider educational sense (Morrow 1986; Messick 1989) derived from theory and experience on the other, as Morrow, Carroll, and the Council of Europe had to do in the late 1970s.

The development of a theory of measurement that can embrace subjectivity rather than trying to minimize its effects means that complex constructs and processes that we do not fully understand do not have to be reduced to numbers in order to be put on a psychometric basis. Objective measurement of subjective judgments has been discovered (Linacre 1989) and appears to work. As mentioned above in the discussion of Pollitt and Hutchinson's study (Pollitt and Hutchinson 1987), the application of a Rasch model cannot validate the posited competence qualities, but it can show how they interrelate with tasks and performances at various levels, and it can draw attention to differences in their interpretation by different types of judges.

3. THE MEASUREMENT ISSUE

A problem with proficiency scales, as with any other test-reporting scales, is that the data tends to be treated—both in statistical analysis and in decision making—as if it were linear, when in fact it is not. Scales of proficiency are certainly not ratio scales (like a thermometer) or interval scales (with equal units of linear measurement); they can at best be regarded as ordinal scales—descriptions ranked in order of difficulty. This is not an argument against scales of proficiency at all; the same thing applies to all test scores. The result of the fact that measurement is not linear is error. When this is associated with a particular test, the error tends not to be noticed until tests or examinations are paired or equated.

A particular problem with scales that have descriptors is the question of deciding the hierarchy between the elements used in the descriptors. If this problem is not solved, the descriptors will systematize error in the subscale concerned—as has been seen to happen with some behaviorally anchored rating scales (Murphy and Constans 1987; Murphy and Pardaffy 1989).

Wilson lists the ways in which learning hierarchies are established: (1) according to supposed psychological characteristics; (2) in logical sequence (but, as Piaget has pointed out, a sequence considered by subject experts to be logical is not necessarily a learning sequence, a point amply demonstrated by SLA research); (3) by order of instruction, which may be arbitrary; and (4) on the basis of a claim to empirical data (Wilson 1989). To these one could add (5) by convention.

In early literature on the ACTFL Guidelines, the claim was made that the guidelines were empirically based, whereas according to critics the base was actually experiential; an order decided by one of the other methods mentioned above was field-tested and "worked." The same can be said of most if not all language proficiency scales (including Eurocentres) and many behavioral scales in work evaluation. In fact, as was mentioned above, recent research using both expert opinion and analysis of scores on tests designed following the guidelines has demonstrated that, at least at the level of detail of novice–intermediate–advanced (i.e., as a threshold scale), the ACTFL descriptors do empirically demonstrate a hierarchy on a linear scale (Dandonoli and Henning 1990; Kenyon and Stansfield 1992). The research used a measurement model called the Rasch model.

The Rasch Model

The Rasch model—a simple, one-parameter model of "item-response theory" (IRT), a branch of "latent-trait theory"—offers a way to calibrate items and persons independently on a truly linear scale in order to establish a hierarchy, or to relate an existing hierarchy to a linear scale. It is called a one-parameter model because it deals only with one parameter—difficulty. There are also two-parameter models taking account of the discrimination of items in addition to difficulty, and three-parameter models taking account of guessing (e.g., in multiple-choice tests). Because the two- and three-parameter models are extremely complex, not very robust, and difficult to work with, the vast majority of IRT applications use the Rasch model. (Good, short overviews of how Rasch can be applied to language teaching are available from Henning 1984; Woods and Baker 1985; and Pollitt and Hutchinson 1987.)

The Rasch model uses a true interval scale of the logarithm of the probability that a person or item will be placed at a certain level. This value is computed on the basis of all the decisions made in the rating or testing. The unit on the scale is called a logit, and because it is a logarithm, zero falls in the middle of the scale, which progresses to infinity on either side. That solves the problem of the "point of origin"—where to start counting—and gives a mathematically true interval scale going from zero to infinity.

Certain reservations were expressed about the educational implications of the Rasch model when personal computers first made it available. These reservations relate to two main points:

— The assumption that the calibrations of an item bank, once established, are true for all time, when in fact curriculum developments over a period of years may mean that some items get "easier" and others "harder" (Goldstein 1981; Tall 1981). This is in fact not a problem confined to the Rasch model, but one pertaining to any standardized scoring system. The Rasch model actually facilitates recalibration or calibration checks to detect value shifts over time.

— The assumption that the calibrations of an item bank developed with one population (e.g., Chinese schoolchildren) can be applied to another (e.g., Scandinavian adults) (Goldstein 1981; Tall 1981). This is a delicate problem, since it requires a value judgment about the point at which a group ceases to be a variant of the same population and becomes a new population. However, the answer can be established by doing an independent analysis of the group in question, then adding the new group to the main data set and comparing results to see if there is a significant problem (Linacre 1992).

Problems have also been known to occur in the calculations of complex algorithms (the operationalization of the model), especially in new programs, but these can easily be dealt with by using two different programs, even two different algorithms, and comparing results (Pollitt 1992).

The Rasch model fits items and persons onto a linear scale by making three assumptions: (1) unidimensionality, (2) local independence, and (3) no guessing. In fact, the model is relatively robust in that it can cope with a certain degree of violation of all three assumptions.

The Rasch requirement for unidimensionality is not to be confused with global proficiency. The claims by Oller were that a "g" factor representing the operationalization of an expectancy grammar was unitary in the same way that IQ represented an unitary concept of intelligence (Oller 1976, 1983). Psychologists no longer hold to the view that intelligence is unitary, any more than most applied linguists would consider language proficiency to be unitary; both claims were partly based in theory and partly an interpretation of the large general factors that tend to be produced by exploratory factor analysis. At the end of a decade of controversy, Skehan and Weir concluded as follows:

> The empirical evidence that has been marshalled in favour of the "unitary competence hypothesis" is open to some doubt and there is a growing body of evidence favouring a divisibility hypothesis. (Weir 1989, p. 5)

> The extreme form of the UCH [unitary competence hypothesis] is now untenable. The Carroll data re-analysis suggests that language proficiency consists of a general factor plus specific factors concerned with aural/oral skills, literacy skills, and then more specific aspects. . . . Bachman and Palmer propose that proficiency is better conceived of as two correlated but separate traits of speaking and reading. (Skehan 1988, p. 213)

However, the issue can be seen in a different way, from the point of view of whether admittedly real distinctions between different skills or aspects of competence are relevant in terms of the information being reported about the abilities of an individual for a particular purpose. Thinking back to the discussion of the nature of proficiency at the beginning of the 1980s, Davies puts it like this:

> The . . . problem is . . . a . . . philosophical one of whether a distinction between a unitary and a non-unitary competence has any meaning. It appears that it is possible to demonstrate from the data we have that either conclusion is correct depending on the type of analysis of the data we use. In other words both the UCH and the non-UCH are "correct" since they reflect different ways of approaching the same issue. They are both right as we can see on the grounds of common sense in that, at some level, there is a unitary language skill, the level at which the distinctions among the performance skills of speaking, writing etc. are unimportant. But at some other level, these very distinctions become very important when we consider issues such as illiteracy and being better at say speaking than at reading. The issue therefore of whether one or the other is correct becomes a non-theoretical issue, while of course remaining very much a practical one. (Davies 1991, pp. 139–40)

Or in Baker's view,

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

To paraphrase Davies and Baker: The fact that people's individual abilities vary substantially across skills, across qualities or aspects of competence, does not alter the fact that it sometimes makes sense just to consider a global summary outcome. This is done, for example, when deciding which class to put a student into in an intensive language program. The classes are organized along one dimension, ability; the ordering is thus unidimensional. One could add that even here, the quality of the information would be improved if one could give feedback as a profile, to assist matching current position to target profile; but the point being made is that in most placement contexts, the prime decision (which class) is based on one dimension (ability).

There are some doubts as to whether it is a "good thing" to report achievement on a global scale in a common framework, rather than as a profile across skills and qualities of performance (aspects of competence) (North et al. 1992). However, that is a separate issue from unidimensionality. The fact that you can have a profile across components assumes that those components point in the same direction, that they operate on the same dimension and preferably share the same unit of measurement or units that can be related to one another. In other words, strictly speaking one could report a profile in scalar form only if the components shared unidimensionality, however integrated or separately they may come in individual cases.

Fortunately, perhaps, even though we generally accept that language proficiency is nonunitary, many other subjects are even less so. Mathematics, for example, changes nature far more radically with the increase in proficiency, and it also has more clearly separate subcategories, which are almost subdisciplines; yet math can be considered unidimensional in the sense that it sometimes makes sense to report one grade (Linacre 1992). Henning, Hudson, and Turner have demonstrated that the common division of language skills—listening, reading, writing (error recognition), grammar, and vocabulary—can be accommodated within the Rasch model (Henning, Hudson, and Turner 1985); Pollitt and Hutchinson's study shows that tasks that are sufficiently distinct sociolinguistically to produce radically different results can also be accommodated by Rasch; and the studies by Hamp-Lyons and Henning and by Pollitt and Hutchinson suggest that qualities of performance reflecting components of competence appear—to raters, at any rate—to be closer to one another than one might suppose, and certainly to pose no problem in terms of unidimensionality. In other words, skills, tasks, and qualities do not appear to be on different dimensions; the elements are different, and people have different shares of competence on each, so that profile reporting makes sense, but the elements appear to share unidimensionality as the term is being used in relation to the Rasch model.

In any case, Rasch analyses produce their own quality control of whether the data is unidimensional—so-called fit statistics—and there are supplementary methods that can be applied in case of doubt. The method Henning, Hudson, and Turner used is called the Bejar method (Bejar 1980). In this method one calibrates all the subtests separately, drawing scatterplots and regression lines and comparing them. In the study by Henning, Hudson, and Turner, all subtests were found to be within 95 percent t values (i.e., unidimensional).

Another method, proposed by Hozayin (1987), is to supplement a Rasch analysis with multidimensional scaling (Kruskal and Wish 1978). Like the Rasch method itself, multidimensional scaling works at the item level (rather than correlation and factor analyses, which work at the test/questionnaire level), and as in the Rasch model, the output comes in the form of a "map":

— Rasch enables one to see where items cluster on the vertical dimension—where they group into a "level."

— Multidimensional scaling maps a graphic-representation distribution of the items in physical space: items that are different appear away from each other, while items that cluster on different parts of the piece of paper belong to similar "categories."

A combination of the two techniques could not only double-check the unidimensionality of the data for the Rasch analysis; it could also give rich descriptive information—arguably more relevant and informative than quantitative methods. The map distribution of the data on a horizontal dimension given by multidimensional scaling (category clusters), coupled with the information about profiles across tasks and across qualities on the vertical dimension from Rasch (levels clusters), could give valuable empirical information to aid decisions about which subscales to use (for tasks or contexts of use, for qualities or degrees of skill), and about whether to use the same categories for all levels or to alter them at a couple of broad thresholds.

The other two Rasch assumptions are not very relevant to the current discussion.

The Rasch requirement for local independence means that items should not be dependent on one another; it is not necessary to get question 3 correct in order to get question 4 correct. People sometimes think it means that one cannot have an "item cluster" or "item bundle," like a series of questions linked to the same passage, or a cloze test. This point is not very relevant to developing scale descriptors, but Theunissen (1987) discusses how the former can actually be coped with, and Hill (1991) has successfully developed a cloze item bank.

Finally, the guessing problem is mostly associated with multiple-choice and true/false items, and is not relevant to descriptor development.

Relevance of Rasch to Framework Development

Wilson lists three ways in which the use of the Rasch model can assist in establishing a learning hierarchy (Wilson 1989):

— The search for item sets (in our case, descriptors) of homogeneous difficulty can inform the reshaping of the definitions of tasks and, in a test-based model, draw attention to weak operationalization of the tasks in the test items.

— Insights can be gained into problems with our theories of learning and instruction (to do with sequencing, obviously, but also in terms of the size of the step up from one objective, task, or level to the next—the existence of thresholds where the rules of the game change).

— It can give a frame of reference for discussing the behavioral meaning of different levels of attainment in a learning sequence (i.e., you can develop descriptors).

To these one could also add that plotting the (ordinal) scale of descriptors onto a (linear, interval) scale removes the imperative people sometimes feel to try to get all the steps on the scale the same size. The difference is that with the "rating scale model" version of the Rasch model (Wright and Masters 1982), available on several Rasch computer programs, it is possible to see how big the steps actually are. This may inform decisions about adding or dropping levels; some scale points may turn out to be small ones, but they may be significant ones that should therefore be kept as they are. It will also aid final graphic presentation to users in the case of a scale having steps of increasing size.

More fundamentally, Rasch can address two essential points made by Thurstone in connection with scales—actually opinion-poll scales, but the point is still valid in discussing a common descriptive framework:

> [The values attached to the scale] must be as free as possible, and preferably entirely free from the actual opinions of individuals or groups. . . .
>
> If the scale is to be regarded as valid, the scale values of the statements

> should not be affected by the opinions of the people who helped to construct it. This may turn out to be a severe test in practice, but the scaling method must stand such a test before it can be accepted as being more than a description of the people who construct the scale. At any rate, to the extent that the present method of scale construction is affected by the opinions of the readers who help sort out the original statements into a scale, to that extent the validity of the scale may be challenged. (Thurstone 1928b, 1928a, cited in Wright and Masters 1982, pp. 5, 15)

Failure to meet this requirement has often caused the anchors in behaviorally anchored rating scales to import and systematize the random error they seek to reduce. An uneasiness about this factor makes people wary about applying to a wider context a scale that has worked perfectly well in a narrower one. This is an argument given by Spolsky against the adaptation of the FSI/ILR scale and Oral Proficiency Interview—used regularly for a specific purpose by a relatively small group of relatively homogeneous raters to rate a relatively predictable type of candidate—into the ACTFL Guidelines and OPI now used to rate teenagers. Problems were even noticed when the Educational Testing Service (ETS) took over the assessment of young Peace Corps candidates from the FSI in the late 1960s (Spolsky 1992).

As was discussed above in connection with rating categories, scale descriptors tend to be developed by specific institutions for a specific purpose, and there seem to be almost as many ways of dividing up categories as there are institutions. As a result, caution is often advised about transferring to one context a scale developed in another. However, it is also fair to say that this caution can be overdone.

A recent study (Hamp-Lyons and Henning 1991), with a very small rating team of three raters, examined a scale of five (increased to seven) separately defined subscales for different categories, originally developed in the ELTS revision project. Though successfully transferred to one new context (a test in Michigan for high-level students), the scale failed to transfer well to another (the ETS Test of Written English, or TWE, for relatively low-level students). It was reported, first, that the subscales were applied differently to the two samples, with different categories of qualities being "salient" in the two groups of students, and, second, that there was a strong halo effect—performance on the different categories was not clearly distinguished. The study also found that the two TWE raters intercorrelated on linguistic accuracy better on the level with which they were familiar, which is hardly surprising. The finding that different categories are salient at different levels, but that overall there is a high intercorrelation between grades for different aspects of competence (halo effect), recalls the findings of Pollitt and Hutchinson (1987). Apart from that, what is being discussed here is what happens when raters are subjected to cognitive overload (here, seven categories by nine levels equals sixty-three boxes on a grid, far too many), given no training, and asked to rate performances at a level with which they are not familiar.

Nevertheless, concern is often expressed about the authorship and generalizability of scales of proficiency. Pienemann and Johnson, for example, consider that "proponents of measures of communicative competence have not, in general, recognised the indirect, relational character of their instruments":

> While proficiency continues to be defined in such terms (*vague, intuitive,* and, most importantly, *relational*—e.g. *effect on listener*) assessment of communicative competence can only be properly interpreted as a mapping of behaviours—that of testers on the one hand and testees on the other. This kind of mapping is complex and multi-factorial, and the constituent behaviours that go to make it up are not amenable to being "untwined" to allow for one to one correlations to be made between parts and the whole, or for individual factors to be weighted consistently in relation to each other. (Pienemann and Johnson 1987, p. 67)

The "untwining" remains a theoretical problem, despite the degree of consensus between the leading models of communicative competence; it has also been suggested that people in practice "untwine" in different ways, and that performance on tasks or language activities may in any case be more appropriate than subcomponents of competence as the organizing principle of a common framework.

A mapping of behaviors, however—of testers on the one hand and testees on the other—is precisely what the Rasch model offers and is precisely what a common framework of levels attempts to do. Indeed, the metaphor of a map is frequently used in discussing the idea of a common framework (Schärer 1992). Different "partners" in evaluation (e.g., learners, teachers, course writers, potential employers) may have valid but different perspectives (Borman 1974; Lance, Teachout, and Donnelly 1992; Schneider and Richterich 1992), and a fundamental aim of a common framework is to provide these perspectives with a common map—or a common calibration of their individual maps.

Such a pragmatic, "relational" mapping process to "enable learners to find their place and assess their progress with reference to a set of defined reference points" (Council of Europe 1992, p. 39) should not be confused with the provision of a theoretical explanatory model of language and of language learning.

Two examples, both from Australia, of relatively simple applications of a Rasch model to the development of scales of language proficiency are as follows:

- Aspects of competence relevant to writing tasks were placed in a hierarchy. Defined assessment scales were developed from them, and then writing samples were calibrated to the scales. The scale and samples worked as a "ladder of developing competence," a framework to guide scoring (Harris, Laan, and Mossenson 1988, cited in Masters 1988, p. 296).

- The results from test tasks written in order to test specific subskills were analyzed and placed on a linear scale. Once on the scale, the tasks were divided into three groups for Level 1, Level 2, and Level 3 on a certificate reporting results on the tests, and then the discrete tasks were edited into short paragraph-length certificate descriptors (Brown et al. 1992).

The Rasch model can also be used in equating the levels of examinations, though there is a certain controversy over using it for "vertical equating"—equating tests at succeeding levels—unless the items are of a similar type. The position is rather confused by the fact that most studies have involved multiple choice, for which the Rasch model is less suited because of guessing (Skaggs and Lissitz 1986, 1988; De Jong 1986). A framework could presumably be developed directly from existing examinations, but the descriptors for the levels and categories would themselves still need to be developed, and that is what is being discussed in this paper.

Subjectivity of Judgments

There are three possible sources of data for estaolishing a hierarchy among descriptor elements, and all three involve subjectivity in judgment:

— designing/identifying tests targeted at the specific skills and subskills involved, and allocating descriptors to levels on the basis of the place in the Rasch hierarchy of the test results;

— appealing to teacher or user judgments, "expert opinion": asking people how far they get their students, or how likely their students are to be able to perform a certain task; and

— rating observed behavior of students.

Using test items targeted at specific skills. In a Swiss context, where evaluation is based almost entirely on teacher assessment rather than on examinations, suitable tests exist only at the point of leaving compulsory education at age sixteen, the so-called Treffpunkte, points of encounter (Vonarburg 1992; Walther 1991), and at the standard achieved at age eighteen by the majority of apprentices through weekly vocational education classes at commercial schools (Dubacher 1989). The Matura examination, taken at age eighteen or nineteen at the end of gymnasium, is a traditional exam organized at a local rather than national level, which would not yield much task-related information.

In any case, there are severe problems in taking test items as operationalizations of particular skills, as Brown and others have done in Australia (Brown et al. 1992; Alderson and Lukmani 1989; Alderson 1988, 1990). Summarizing studies on native-speaker-teacher judgment of items testing reading subskills, Alderson concludes:

> i) Judges are unable to agree as to what an item is testing; ii) Judges are unable to agree upon the assigning of a particular skill to a particular test item; iii) Judges are unable to agree about the level of a particular skill or a particular

> item; iv) There appears to be a lack of relationship between item statistics and what an item is claimed to be testing. (Alderson 1988)

In view of the difficulty of writing test items to assess a given aspect of competence (a difficulty encountered by the Canadian Development of Bilingual Proficiency Project—Harley et al. 1990), and in view of the inconsistency of teachers'/testers' judgments on what items are actually testing (Alderson 1988), there appear to be severe difficulties in using test items targeted at specific skills to collect the data. Brown and others used this approach but validated the test items they used through an independent analysis—but such an approach implies a doctoral dissertation per test.

Appealing to "expert opinion." The second approach also seems highly questionable, because "expert opinion" of hypothetical difficulty has also been shown to be very erratic. Alderson found that experienced and inexperienced non-native-speaker testers in Sri Lanka were unable to predict the difficulty of test items in any clearly discernible pattern (Alderson 1990). In a separate study by the National Foreign Language Center, a hierarchy of difficulty for an Arabic test derived through a Rasch model analysis of Egyptian native-speaker teachers' estimation of difficulty proved to show little discernible relationship to the performance of American students on the test (Lambert 1992).

In another project, Alderson asked non-native teachers to judge cutoff scores for grades on an examination:

> Think of pupils you have taught. Think of people who you consider to be just barely a pass at O-Level English Language. Call them "Bare Pass." Think of other pupils you would consider to be only just barely a credit at O-Level English Language. Call them "Bare Credit." Think of a third group of pupils who you consider to be just barely a distinction in O-Level English Language. Call them "Bare Distinction." (Alderson 1990)

Teachers were asked to judge the cutoff scores for the three groups in three ways: (1) paper level (what mark would be achieved on each of the two papers—i.e., the grade cutoff point); (2) item level (what percentage would get each question correct, or what score they would get on a writing question); and (3) overall (what overall score out of 100 would be the cutoff point for the examination as a whole). The paper-level and item-level estimates were unsatisfactory—the one substantially overestimating the traditional proportion passing, the other substantially underestimating it. The overall judgment was close to the traditional cutoffs but underestimated the mark necessary to get a distinction. Alderson's conclusion is that when we think of problems with judgments in language assessment, we should not just focus on raters. All judgments in the testing process need to be corroborated before they can be accepted as valid (Alderson 1990).

Rating observed behavior. Since the first two methods are problematic, and since it proves to be easier to deal with subjectivity when it is a rating of something observed rather than an abstract impression, this suggests following an observation-and-rating strategy. In the following discussion the terms *rater* and *judge* are used synonymously; some writers prefer the one, some the other.

Subjectivity in Rating Observed Behavior

This third method, rating behavior, can be used in two ways, as was discussed under behaviorally based rating scales:

— Observing and rating the behavior of a number of students in a class over a period of time, using a checklist of tasks, quality statements, and summary statements. As with behavioral observation scales (BOS) in work evaluation, this would be a mixture of direct observation, retrospective observation, and expected behavior.

— Rating samples of behavior representative of the range of levels in the system, including oral samples of the two examinations mentioned earlier (Treffpunkte, commercial schools), on an assessment scale of defined bands related to the quality statements in the BOS checklists (behavior summary scale, BSS).

The problems of interrater reliability (consistency between raters) associated with the above are well known, and they are the reason some people have pursued supposedly "objective" strategies like error counts (for which there is even a version of the Rasch model, called the Poisson count).

The problem was recognized at the end of the nineteenth century. Edgeworth estimated the degree of chance in public examinations to be between one-third and two-thirds, and Ruggles noted that the amount of variance between judges was as great as that between candidates (Edgeworth 1890 and Ruggles 1911, cited in Linacre 1989, pp. 10–11). The American National Board of Medical Education dropped subjective assessment after studies demonstrated interrater reliability of only 0.25 percent (Hubbard 1971, pp. 93–99, cited in Raymond, Webb, and Houston 1991, p. 101). Not much progress has been made in that area in the health professions. Cason and Cason recently demonstrated that 35 percent of variance was due to the strictness of the rater, and only 40 percent to ability (Cason and Cason 1984).

A recent review of interrater-reliability studies reports some findings in the 0.70s and 0.80s but stated that the majority were around the 0.40s to low 0.60s—meaning that the estimated variance accounted for by ability is only 20 percent to 40 percent (the square of the correlation) (Muzzin and Hart 1985, cited in Raymond, Webb, and Houston 1991, p. 101). Borman is reported to have conducted a very carefully controlled experiment with a rigorously constructed scale, well-chosen samples, trained raters, and laboratory conditions, and to have achieved only 0.80, or 64 percent (cited in Gruenfeld 1981, p. 12, cited in Linacre 1989, p. 10).

The general response has been a trend toward defined descriptors and a focus on training. Jason gives typical advice, suggesting that the scales be made as clear as possible by

— refining the aspects to be rated (defining the trait, adding concrete descriptors; defining the qualities to be looked for, providing descriptions of the behavior to be observed);

— refining response categories; and

— training the raters.

In this way, Jason claimed, scale reliabilities of 0.86 to 0.93 were obtainable (Jason 1962, cited in Wolf 1988).

Language testing literature reporting results from experiments using carefully developed, defined scales and trained raters often seems to give levels of interrater reliability higher than what appears to be the average in work performance. For example, in a recent study on the ACTFL Guidelines, Dandonoli and Henning report mean interrater reliabilities between ACTFL-trained examiners and "naive native speakers" of between 0.929 and 0.857 (Dandonoli and Henning 1990). Milanovic and others report mean interrater reliabilities of 0.93 during the development of the rating scales for the new Cambridge Assessment of Spoken English examination (Milanovic et al. 1992). However, these are obviously "lab" results; the average is clearly a lot lower.

People have tried to reduce the element of chance in ratings by systematizing procedures, providing rigorous training, and using various raw score conversions. None of these methods has worked very well, however, because they treat the data as if it were linear (Engelhard 1991) and tend to use aggregate scores even though the actual problem is thought to lie with the individual rater (Saal, Downey, and Lahey 1980). Many writers would not in any case accept that interrater reliability is in itself a viable goal, following the argument that if a group of raters agree entirely, the one thing certain is that the agreed rating is wrong; agreement may be an agreed bias rather than a valid measure (Saal, Downey, and Lahey 1980). This point can be taken further, to say that the reliability of a rating scale, however established, says nothing about its validity, since the reliability may be just consistent bias and is not synonymous with rating accuracy (Wherry 1952).

The situation is further complicated by the fact that we do not really know how people rate, any more than we know how people learn. According to Einhorn, people may differ radically as to how they identify information, organize it into clusters or dimensions, and then weigh and measure it. He concludes that "in a highly probabilistic world, there may be many routes to the same goal" (Einhorn 1974).

The classic rater errors are halo effect (transferring judgments from a global impression to categories, or between categories); central tendency (not using the top and bottom of the scale, or tending to home in on a neutral category on a questionnaire); and variation in severity/leniency. Training in the work performance evaluation field tends to concentrate on these points—that is to say, on changing rater behavior. There are indications that these problems can be reduced by video workshop training (Cooper 1981, p. 233, in relation to halo reduction; Ivancevich 1979, in relation to halo and severity/leniency) and by related diary keeping prior to assessment (Bernadin and Walter 1977, in relation to leniency). Paradoxically, however, "successful" training to change rater behavior does not necessarily increase rater accuracy or interrater reliability, and it can in fact reduce both (Bernandin and Pence 1980; Borman 1979). Finally, the effects of training appear to diminish over time (Ivancevich 1979; Cooper 1981, citing Warmke and Billings 1979; Bernadin 1978).

These findings may reflect the fact that at least two of the classic errors—halo effect/holistic rating and severity/leniency—are caused by personal characteristics of the rater and are extremely resistant to training (Cooper 1981, pp. 226–39, on halo; Linacre 1989, on severity). To be "effective," therefore, training concentrating on these errors presumably needs to destabilize the rater's natural approach, disorientating the rater in the process—which may be what causes the kind of loss of reliability and accuracy reported (Bernandin and Pence 1980; Borman 1979).

Training aimed at changing rater behavior in relation to classic errors may in fact be totally misconceived. After a review of studies bearing upon the relationship between rater accuracy and halo effect, Cooper reaches a devastating conclusion: "The best available estimates suggest that halo and accuracy share a median of 8% of the variance, but the direction is opposite to the prevailing assumption—that is, higher halo and higher accuracy modestly covaried" (Cooper 1981, p. 239).

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

Borman suggests that, rather than focusing on classic rating errors, training to improve interrater reliability should seek a "common nomenclature" like a "frame of reference" for defining effectiveness levels, as well as standardized observation and agreed-upon weightings of qualities. He suggests that training to increase accuracy requires that these agreed-upon effectiveness levels and the weights attached to different factors be "correct" and "uncontaminated" by irrelevant aspects, which implies that they should be objective measures of subjective judgments (Borman 1979).

The suggestions from Einhorn, Bernadin and Pence, Cooper, Cason and Cason, and Borman seem to confirm a lot of empirical and anecdotal evidence from language testing, and directions in which the assessment of oral interaction seems to be moving. For example, the Eurocentres approach to the assessment of spoken performance in group interaction uses two assessors: one who knows the class (high sensitivity), and one who knows the whole range of the level (low sensitivity) whether or not he or she knows the class. The procedure does not try to prevent halo effect, and although grades are allocated to each rating factor, this is done after giving an initial holistic impression based on a defined scale, and the final overall grade is not arrived at by arithmetic (echoing the old FSI system). The procedure in fact encourages both a holistic and an analytic approach and a synthesis of the results, with negotiation over grades between the two assessors as a final step to adjust for severity. This approach tries to accommodate and allow for the chance factor of severity; what is important is that people be consistent, true to themselves. If a rater is too severe, this is dealt with in the negotiation with access to detailed descriptors and resort to a higher authority to moderate if necessary (North 1991, 1992a). Video examples that have been rated and re-rated after a twelve-month interval are available for training and as a reference of last resort.

As evidence mounts that judge severity is relatively impervious to training and that people rate in different ways, such negotiated approaches are gaining ground (Porter 1991).

The Linacre Many-Faceted Rasch Model

We have considered Borman's requirement for "correct," "uncontaminated" effectiveness levels and factor weightings—in effect, for objective measures of subjective judgments (Borman 1979)—and Thurstone's requirement that the values attached to the scale be independent of the opinions of the persons who helped write it (Thurstone 1928a, 1928b, cited in Wright and Masters 1982). Such requirements suggest that in formulating an approach to standard setting and evaluation that embraces and values the subjectivity of informed professional judgment, it would be logical to seek some means of identifying and if necessary adjusting for that subjectivity in the actual measure, particularly during the process of defining "effectiveness levels" and evaluating the video examples that will be the main standardization instruments for training. The Lincare many-faceted Rasch model offers a way to do this.

The Linacre Rasch model differs from previous Rasch models, and from previous treatments of interrater reliability, in that it takes the chances of having a severe or lenient rater into account and estimates them in the same way it estimates the chances of having a hard or easy question. It gives item-free, person-free, judge-free measurement, as well as (if desired) information about the size of the steps on the rating scale. The item, person, judge, and scale step are all "facets" of the evaluation. Other facets such as the rating occasion can also be added. The model allows the interaction of these different facets to be studied and appropriate adjustments to be made to arrive at a fairer, more accurate judgment.

The approach is mainly concerned with severity (the judge); interestingly, however, this focus on how facets interact also reflects the state of the art on research into the halo effect. Murphy and Anhalt propose that rather than being a rater-related issue, as it was interpreted by Cooper (1981), "halo error may reflect a wide variety of influences including the rater, the ratees, and the specific behaviour that is being evaluated at a given point in time" (Murphy and Anhalt 1992, p. 499). In the terminology of studies that have used the Linacre model, rater = judge; ratee = person; specific behavior = item (or item on a task); and given point in time = rating occasion. This latest study echoes earlier complaints from Landy and Farr and from Saal, Downey, and Lahey that the concentration on rater behavior fails to acknowledge the complex interaction among raters, ratees, traits, methods, times, and so on (Landy and Farr 1980; Saal, Downey, and Lahey 1980).

Like all item-response models, the Linacre model uses a linking network of anchors, so it requires a data-collection design that gives a linked network. Unlike many other such models, it can cope with incomplete data and is very economical in the way the network is constructed.

Application of the Many-Faceted Model to Framework Development

In establishing a common framework of descriptors and samples that may be used by future judges to report achievement, it could be particularly important to accomplish the following aims:

- Involve people representative of the user populations in the process of development (Smith and Kendall's suggestion—1963).
- Incorporate description elements from "feeder systems" that will later help identify links between those feeder systems and the framework (anchoring).
- Separate the parameters—person, task, judge, groups of judges (Thurstone's requirement—1928b).
- Establish how different groups of users interpret different kinds of descriptors (Borman 1979; Schneider and Richterich 1992).
- Do rigorous item analysis during framework construction to identify false assumptions before they systematize the error they are designed to reduce—a common failing with behaviorally anchored rating scales (Landy and Farr 1983).

- Establish the plausibility of the model, the way real ratings relate to or "fit" the model. Can people use it?
- Establish that the elements of the system (tasks, quality statements, holistic generalizations, actual performances related to those) demonstrate coherence—that they can be measured on one dimension.
- Establish the degree to which levels of performance can be distinguished—the number of level strata that can be deduced from the way people rate.
- Check the existence and real relative size of the steps on the proposed scale, and compare them with the empirically determined level strata in the data. Discover how the order and segmentation in the hierarchy relate to those expected, and act accordingly.
- Work interactively incorporating data, making revisions until a satisfactory hierarchy is established.
- Do rigorous quality control of the wording of the resultant draft descriptors.

All these aims can be achieved with the Linacre version of the Rasch model in the program FACETS, which is the latest in a series of Rasch programs developed in the MESA laboratory at the University of Chicago. The only restricting requirement is that the data matrix have adequate links in it.

Relevant FACETS Studies

A number of studies have used the FACETS model since its appearance in 1988, and the most significant are summarized below.

Myford assessed how the acting ability of high school students was interpreted by different groups of judges: experts, theater buffs, and novices. She used a six-point Likert scale in a questionnaire of thirty-six qualities, each identified by a label (Myford 1991). For experts, buffs, and novices, one could imagine grouping by geographical region, language (native/non-native speaker), or role (teacher/student/employer).

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

In a physiotherapy study, Fisher looked at how patients performed four steps (items) of eight composite exercises with objects (tasks), with performance being rated in terms of four categories (quality criteria) identified with a label on a four-point rating scale (Fisher 1991).

Kenyon and Stansfield investigated how three groups from a total of 402 teachers (bilingual education teachers, French language teachers, and Spanish language teachers) ranked thirty-eight tasks from the ACTFL Guidelines; the teachers were asked to rate each task on a five-point Likert scale, indicating whether a teacher of French/Spanish in Texas should be able to perform that particular task. In a second study, the same teachers were asked to rate fifteen to seventeen audiotape extracts from interviews previously rated on the ACTFL scale by ACTFL raters as yes or no in answer to the question whether this person had a sufficient level of French/Spanish to teach in a Texas classroom. The answers were used to determine a hierarchy among the tasks and among the extracts—a hierarchy that coincided very closely with that intended in the ACTFL Guidelines, with the exception of four tasks having a high nonlinguistic or personal dimension (Kenyon and Stansfield 1992).

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

Number of Levels/Points/Bands

Another measurement question, this one concerned with the "vertical" axis of the framework, is the number of levels to adopt (North 1992b). In an approach aimed at improving transparency for the various partners in the system, there is some argument for having a scale of defined bands or steps rather than a continuous numerical reporting scale (e.g., 1 to 1,000). This philosophical and educational argument has received psychometric support from a study by McKelvie showing that a continuous numerical scale offers no advantages in terms of reliability or validity (McKelvie 1978).

The question, then, becomes how many defined steps to have. The argument for fewer, broader bands is primarily a psychometric one, and the argument for more, smaller steps is primarily an educational one (North 1992b). As the series of FACETS studies just referred to illustrates, there is a tendency to use rating scales with between four and six steps.

In the majority of studies investigating the optimal number of steps, the optimal number has been related to the reliability of the scale (McKelvie 1978):

— Empirical research from the 1950s suggests that maximum reliability is reached with five steps; this reliability remains constant up to nine steps and tails off with either three or eleven (Bendig 1953, 1954a, 1954b, cited 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 report stable reliability from two to nineteen categories (Matell and Jacoby 1971), but other studies suggest no increase in reliability above six categories (McKelvie 1978). Miller summarized his findings with the rule of thumb "seven, plus or minus two," pointing out that psychologists even then had long been using seven-point scales on intuitive grounds (Miller 1956). McKelvie concludes by recommending five or six (McKelvie 1978).

- Pollitt explains the relationship between reliability and the decision capability from any test (or rating). A reliability of 0.96 is needed for ten bands, 0.90 for six bands, and 0.80 for four bands (Pollitt 1991). Considering the points made earlier about commonly reported levels of interrater reliability, this suggests that five levels is optimistic for operational systems.

- An approach used in several studies focusing on the standard deviation of discrimination—studies suggesting that information gains were rapid up to ten or twelve categories, and that small gains were made up to forty-four categories—has been demolished on methodological grounds. The studies in this area suggest that there is no advantage to be gained from using more than twelve categories (McKelvie 1978).

- Andrich and Masters conclude that more steps will increase precision up to the number of decisions the person can cope with—a probable maximum of nine (Andrich and Masters 1988).

While to some extent contradictory, the evidence suggests that five to six steps is a safe number, but that systems of up to nine or twelve can function quite adequately.

In connection with scales of language proficiency, it is important to distinguish between the number of decisions any one person is making and the number of levels existing in the framework as a whole. The IELTS test has been criticized (e.g., by Hamp-Lyons and Henning 1991) for having nine levels, which is known to be more than necessary and more than raters can handle. The Cambridge Assessment of Spoken English test reduced the number of bands from nine to six on empirical evidence (Milanovic et al. 1992).

In Eurocentres we have a ten-level system, though for some purposes, including placement interviews, these levels are grouped into five categories: beginner (1), elementary (2–3), intermediate (4–5), upper intermediate (6–7), and advanced (8–10). In progress and exit assessment, on the other hand, since students are in classes by level, it is extremely rare to come across a range of more than three performance levels in a class, even at the end of a course of three months. During progress and exit assessment, we therefore also use "plus levels"—giving twenty points in the framework as a whole. But in practice, no rater has to decide between more than five or six points (e.g., 3, 3+, 4, 4+, 5, or 5+), because there will be only a limited range of level in the class, and the first step of the assessment procedure is to establish what that range is, in order to concentrate on the right part of the scale.

This important distinction tends to get missed in discussion of this issue, particularly since most of the literature relates to what are in effect questionnaires rather than scales of proficiency.

There seems to be a tendency for the number of levels/bands/points/

grades to increase as a scale develops from being used as a rating scale to being used as a framework. The FSI/ILR and ACTFL developments are a case in point. The original FSI scale was a simple 0–5 six-point graphic rating scale. By the time of Wilds's publication of the scale (Wilds 1975), "plus levels" were being used between the defined levels. By 1983 the "plus levels" in what was now called the ILR scale were fully defined, giving eleven defined levels, and the new ACTFL Guidelines—also derived from the FSI scale, and intimately related to the ILR scale—had expanded the lower three levels (0–2) so that Level 0 equaled novice low and mid; Level 0+, novice high; Level 1, intermediate low and mid; Level 1+, intermediate high; Level 2, advanced; and Level 2+, advanced plus. Anything above that (3–5) was superior.

In a large-scale (N = 231) experiment with minimally trained raters, Meredith has suggested taking this process a stage further by adding what is in effect a mixed standard scale (MSS) to the nine-band ACTFL scale in which raters assigned zero for an average or middle performance, plus for an above-average one, and minus for a below-average one (Meredith 1990). Using a multiple regression model relating band scores to previous learning experience, Meredith was seeking to determine whether a numerical version of a modified oral proficiency scale would be more feasible for research. The modification was the MSS rating mentioned above; the numerical version was a set of alternative graduated-interval (rather than equal-interval) scales. All the modified versions showed an increase in the accuracy of the correlation. What is relevant to the present discussion is that just adding the plus and minus scores to the ACTFL grades *increased* the accuracy of the regression by 4 percent. This may not seem like much. But it is unusual in that, as discussed at the beginning of this section, introducing finer-level distinctions is usually said to *reduce* correlations and reliability; and it represents an improvement nearly as great as the 5 percent increase in reliability claimed for the addition of behavioral descriptors to numerical or graphic scales.

Griffin reports on a rating scale used for teacher observation of student behavior during the field-trialing of band descriptors for a new reading scale. That is to say, each level or band on the scale of proficiency can be subdivided with a four-point 0–3 rating scale. The wording of this rating scale takes account of partial attainment in what is seen as a developmental process; it tries to encourage explicit observation and close comparison with the criterion descriptor, rather than the norm-referencing invited by Meredith's approach:

> 3. If the student *has established the behaviour pattern* and consistently exhibits all or most of the behaviour described in the band.
>
> 2. If the student *is developing the behaviour pattern* such that some but not all of the behaviour described for a band is exhibited. . . .
>
> 1. If the student *is beginning to show signs of the behaviour pattern* of a band level in that only a little of the pattern is shown. . . .
>
> 0. If the student *shows none of the behaviour pattern* of a band. (Griffin 1989)

This rating scale was a research tool used to give input to a Rasch rating scale model analysis. Such a technique could also be used to help identify the match between the bands on the scale and attainment in school years—perhaps to see if the definitions need adjustment up or down to realize the intentions of the scale.

Teachers in Eurocentres have sometimes expressed a wish to discriminate between performances in the same band in order to be better able to demonstrate progress—or to norm-reference, to rank students. As was mentioned at the beginning of this paper, several writers have pointed out that although scales of proficiency are the logical conclusion in the development of criterion-referencing away from mastery/nonmastery knowledge of discrete items and toward the recognition that performance takes place along a continuum of developing mastery in the trait, scales of proficiency also blur the distinction between criterion- and norm-referencing.

Attaching a rating scale to band levels on a scale might be a way to offer more delicacy, more steps to climb through. A case can certainly be made for rewarding excellence at a low level; a case can be made for judging performance in terms of what it is reasonable to expect from a child at a certain stage of development. One could even use the pass-credit-distinction grades, to avoid more numbers. As with Griffin's development tool, it could be done in a way that also seeks to put the emphasis on the stage of the developmental process, the stability of the behavior in the student's variable interlanguage (Tarone 1983; Ellis 1986, 1987). For example:

> 4. *Distinction:* Very stable performance at this level, even in very relaxed or stressful situations. Performance often peaks above this defined level, but the student cannot sustain a performance at the next level.
>
> 3. *Credit:* Stable performance at this level. Can in fact peak above the standard defined for this level in optimum conditions.
>
> 2. *Good pass:* Stable performance at this level in neutral conditions. May sometimes backslide below the standard defined for this level in very relaxed, tiring, or stressful situations.
>
> 1. *Pass:* Performance is usually sustained above the borderline for this level but is not yet fully stable. May very well backslide below the standard on occasion.

One does need to remember, however, that such an approach brings with it the danger that teachers might decide that their class is Level 4 because it is doing the book for Level 4, then rank the class and give out such named grades for Level 4 (pass, credit, etc.) on the basis of class test results or "effort" rather than consulting the criterion.

Whether one decides to have finer levels on the main scale, consulting with teachers about how many levels they feel would be adequate to show progress and in effect stringing a series of miniscales into an overall scale (North 1992c); or whether one decides to have broader levels, like the Australian Second Language Proficiency Ratings, and match teacher grades to them (Ingram and

Wylie 1989); or whether one attaches a rating scale to the bands of a proficiency scale (Griffin 1989)—in any case, one of the things a Rasch analysis will do is to identify the degree of discrimination, the number of level strata in the data, the number of categories the data can bear, the number of decisions people display an ability to make. In a Rasch analysis the figure that gives this information is the separability statistic, which is the Rasch equivalent of a KR 20 reliability figure (Wright and Masters 1982). It would thus be possible to base the decision about the number of levels in a descriptive framework directly on a measurement model.

4. CONCLUSION

This paper has looked at some of the theoretical issues underlying the development of scales of language proficiency, starting by listing the problems that have been identified with existing scales and the reasons why people perceive a need to develop them. It was pointed out that a scale of proficiency must be an operational model, and while it can take account of developments in theoretical models of language competence, it may therefore adopt descriptive categories that relate to observable behaviors rather than the ability traits thought to underlie them. A model is in any case a simplification rather than a replication of reality, and its role is to help people organize and understand that reality. To do so, it should be couched in metalanguage that is accessible and comprehensible to the people who will use it; they should preferably be consulted in the process.

The main different schools of presenting information through behaviorally based scaling in work evaluation were discussed, and it was pointed out that the difference among the formats—behaviorally anchored rating scales, behavior observation scales, and behavior summary scales—was largely one of presentation. The main problems found with such scales have been due to inadequate item analysis and insufficiently rigorous determination of the level of the anchors (tasks), with the effect of systematizing rather than eliminating error. The anchors tended themselves not to be anchored in a measurement model.

Experience suggests that conventional quantitative (test-level) methods for empirically determining the construct validity of a scale, or of the descriptive model on which it is based, may prove less enlightening than the "mapping" (item-level) techniques offered by a many-faceted Rasch model, possibly augmented by multidimensional scaling. The interaction between qualities and sociolinguistic tasks at different levels on the ability continuum may be more significant than the actual competence model on which the qualities are based.

Information about tasks and information about degrees of skill in performance qualities are both necessary, and they are best kept separate because (1) they tend to be used for different purposes in the decision-making part of the system (constructor-oriented, assessor-oriented) and come together only in reporting results (user-oriented); (2) unless they are separately defined, the interaction between them at different levels cannot be adequately evaluated; and (3) separating the two gives more flexibility, which may be important since different institutions use different qualities as criteria, and some raters rate more holistically than others.

The Rasch model offers a methodology to

- do item analysis—identify which descriptors and bits of descriptors work;
- establish an empirical hierarchy—in our case, identify how different descriptors and bits of descriptors are interpreted in terms of level; and
- establish the relationship among the outcomes of different educational sectors on a genuinely linear, common, defined scale.

The many-faceted FACETS Rasch approach has the added advantages of offering a way to

- adjust decisions about the level of descriptors or samples of performance, to take account of the severity/leniency of judges (i.e., get more accurate measures); and
- involve different partners (teachers, students, potential employers), sectors, and regions, identify differences of interpretation, and inform decisions on whether and how to adjust measures for them.

A particular attraction of a Rasch approach would be that the item analysis, scale validation, sample calibration, and so on would all be part of the same databank, simplifying the development process. This databank could be continually expanded in a project spiraling outward in a series of phases while being able to give a concrete outcome with final calibrations at the conclusion of each defined phase.

In this way one could meet the ultimate test of scale generalizability, by separating the values given to items that make up the descriptors in the scale from the opinion of the people who were involved in its development. This would not remove subjectivity completely; but it would mean that the values used were those of a wide consensus, reflecting different perspectives, just as the categories used were based on the consensus on theoretical and operational models of language competence taken into account when constructing the scale.

Such an approach to the development of a common descriptive framework would provide an empirical psychometric base to support it—the framework would be rooted in a measurement theory—which would help to increase the currency of the eventual reporting instrument.

REFERENCES

ACTFL (American Council on the Teaching of Foreign Languages). 1986. Proficiency Guidelines.

——. 1989. Oral Proficiency Interview evaluation grid.

Alderson, J. C. 1988. "Testing Reading Comprehension Skills." Paper presented at the Sixth Colloquium on Research in Reading in a Second Language, TESOL, Chicago. Cited in Alderson 1990.

——. 1989. Personal communication.

——. 1990. "Judgements in Language Testing, Version Three." Paper presented at the Ninth Meeting of the World Congress of Applied Linguistics, Thessalonica, Greece, April 15–21.

——. 1991a. "Bands and Scores." In Alderson and North 1991, pp. 71–86.

——. 1991b. "Washback or Backwash: Request for Contact/Cooperation." *Language Testing Update* 10:72–73.

Alderson, J. C., and Y. Lukmani. 1989. "Cognition and Reading: Cognitive Levels as Embodied in Test Questions." *Journal of Reading in a Foreign Language* 5/2:255–70.

Alderson, J. C., and B. North, eds. 1991. *Language Testing in the 1990s*. Modern English Publications/British Council. London: Macmillan.

Allen, P.; J. Cummins; R. Mougeon; and M. Swain. 1983. *The Development of Bilingual Proficiency: Second Year Report and Appendices*. Toronto: Modern Language Centre OISE. Cited in Bachman 1990a.

Andrich, D., and G. Masters. 1988. "Rating Scale Analysis." In Keeves 1988, pp. 297–303.

Bachman, L. 1987/88. "Problems in Examining the Validity of the ACTFL Oral Interview." In *Proceedings of the Symposium on the Evaluation of Foreign Language Proficiency*, ed. A. Valdman, pp. 29–43. Bloomington: Committee for Research and Development in Language Study, Indiana University, 1987. Rpt. in *Studies in Second Language Acquisition* 10/2 (1988): 149–64.

——. 1989. "The Development and Use of Criterion-referenced Tests of Language Proficiency in Language Programme Evaluation." In *The Second Language Curriculum*, ed. R. K. Johnson, pp. 242–58. Cambridge: Cambridge University Press.

——. 1990a. "Constructing Measures and Measuring Constructs." In Harley et al. 1990, pp. 26–38.

——. 1990b. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L., and A. Palmer. 1981. "A Multitrait Multimethod Investigation into the Construct Validity of Six Tests of Speaking and Reading." In Palmer, Groot, and Trosper 1981, pp. 149–65.

——. 1982. "The Construct Validation of Some Components of Communicative Proficiency." *TESOL Quarterly* 16:449–65.

——. 1983. "The Construct Validity of the FSI Oral Interview." *Language Learning* 31/1·67–86.

——. 1989. "The Construct Validation of Self-Ratings of Communicative Language Ability." *Language Testing* 6/1:14–29.

Bachman, L., and S. Savignon. 1986. "The Evaluation of Communicative Language Proficiency: A Critique of the ACTFL Oral Interview." *Modern Language Journal* 70:380–90.

Baker, D. 1989. *Language Testing: A Critical Survey and Practical Guide*. London: Edward Arnold.

Barnes, D., and F. Todd. 1977. *Communication and Learning in Small Groups*. London: Routledge and Kegan Paul.

Barnwell, D. 1991. "Proficiency Testing and the Schools." *Hispania* 74/1:187–89.

Beebe, L. 1983. "Risk-taking and the Language Learner." In *Classroom-oriented Research in Second Language Acquisition*, ed. H. Seliger and M. Long, pp. 39–66. Rowley, Mass.: Newbury House. Cited in Brindley 1991.

Bejar, I. 1980. "A Procedure for Investigating the Unidimensionality of Achievement Tests Based on Item Parameter Estimates." *Journal of Educational Measurement* 17/4:283–96.

Bendig, A. W. 1953. "The Reliability of Self-Ratings as a Function of the Amount of Verbal Anchoring and of the Number of Categories on the Scale." *Journal of Applied Psychology* 37:38–41. Cited in Landy and Farr 1983.

——. 1954a. "Reliability and Number of Rating Scale Categories." *Journal of Applied Psychology* 38:38–40. Cited in Landy and Farr 1983.

——. 1954b. "Reliability of Short Rating Scales and the Heterogeneity of the Rated Stimuli." *Journal of Applied Psychology* 38:167–70. Cited in Landy and Farr 1983.

Berk, R. A. 1988. "Criterion-referenced Tests." In Keeves 1988, pp. 364–70.

Bernadin, H. J. 1978. "Effects of Rater Training on Leniency and Halo Effects in Student Ratings of Instructors." *Journal of Applied Psychology* 63:301–8. Cited in Cooper 1981.

Bernadin, H. J., and C. Pence. 1980. "Effects of Rater Training: Creating New Response Sets and Decreasing Accuracy." *Journal of Applied Psychology* 65:60–66.

Bernadin, H. J., and P. C. Smith. 1981. "A Clarification of Some Issues Regarding the Development and Use of Behaviorally Anchored Rating Scales (BARS)." *Journal of Applied Psychology* 66/4:458–63.

Bernadin, H. J., and C. S. Walter. 1977. "Effects of Rater Training and Diary Keeping on Psychometric Error in Ratings." *Journal of Applied Psychology* 62:64–69. Cited in Borman 1979.

Bialystok, E. 1982. "On the Relationship between Knowing and Using Linguistic Forms." *Applied Linguistics* 3/3:101–6.

——. 1986. "Some Evidence for the Integrity and Interaction of Two Knowledge Sources." In *New Dimensions in Second Language Acquisition Research*, ed. R. W. Andersen. Rowley, Mass.: Newbury House.

Borman, W. C. 1974. "The Rating of Individuals in Organizations: An Alternative Approach." *Organizational Behavior and Human Performance* 12:105–24.

———. 1979. "Format and Training Effects on Rating Accuracy and Rater Errors." *Journal of Applied Psychology* 64:410–21.

———. 1986. "Behavior-based Rating Scales." In *Performance Assessment: Methods and Applications*, ed. R. Berk, pp. 100–120. Baltimore: Johns Hopkins University Press.

Borman, W. C., and M. D. Dunnette. 1975. "Behavior-based versus Trait-oriented Performance Ratings: An Empirical Study." *Journal of Applied Psychology* 60:561–65.

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

———. 1991. "Defining Language Ability: The Criteria for Criteria." Unpublished paper.

Brown, A., et al. 1992. "Mapping Abilities and Skills Levels Using Rasch Techniques." Paper presented at the Fourteenth Language Testing Research Colloquium, Vancouver, February 27–March 1.

Burton, D. 1980. *Dialogue and Discourse: A Sociolinguistic Approach to Modern Drama Dialogue and Naturally Occurring Conversation*. London: Routledge and Kegan Paul.

Byrnes, H. 1987. "Second Language Acquisition: Insights from a Proficiency Orientation." In Byrnes and Canale 1987, pp. 107–32.

———. 1989. "Evidence for Discourse Competence in the Oral Proficiency Interview." *Applied Language Learning* 1/1:1–13.

Byrnes, H., and M. Canale, eds. 1987. *Defining and Developing Proficiency: Guidelines, Implementations, and Concepts*. Lincolnwood, Ill.: National Textbook Company.

Campbell, D. T., and D. W. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 56:81–105. Cited in Bachman and Palmer 1982; Dandonoli and Henning 1990.

Canale, M. 1983. "On Some Dimensions of Language Proficiency." In Oller 1983, pp. 333–42.

Canale, M., and M. Swain. 1980. "Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing." *Applied Linguistics* 1/1:1–47.

———. 1981. "A Theoretical Framework for Communicative Competence." In Palmer, Groot, and Trosper 1981, pp. 31–36.

Carroll, B. J. 1978. *Specifications for an English Language Testing Service*. London: British Council.

———. 1980. *Testing Communicative Performance*. Oxford: Pergamon.

Carroll, B. J., and P. J. Hall. 1985. *Make Your Own Language Tests*. Oxford: Pergamon.

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

Cason, G. J., and C. L. Cason. 1984. "A Deterministic Theory of Clinical Performance Rating." *Evaluation and the Health Professions* 7:221–47.

Champney, H. 1941. "The Measurement of Parent Behavior." *Child Development* 12:131–66.

Child, J. R.; R. Clifford; and P. Lowe. 1991. "Proficiency and Performance Testing." Unpublished paper.

Clark, J. L. D., and R. T. Clifford. 1988. "The FSI/ILR/ACTFL Proficiency Scales and Testing Techniques: Development, Current Status, and Needed Research." *Studies in Second Language Acquisition* 10/2:129–48.

Clark, J. L. D., and J. A. Lett. 1988. "A Research Agenda." In *Second Language Proficiency Assessment: Current Issues*, ed. P. Lowe, Jr., and C. W. Stansfield, pp. 53–82. Englewood Cliffs, N.J.: Prentice-Hall Regents.

Clifford, R. T. 1980. "Foreign Service Institute Factor Scores and Global Ratings." In *Measuring Spoken Language Proficiency*, ed. J. R. Frith. Washington, D.C.: Georgetown University Press.

Cooper, W. H. 1981. "Ubiquitous Halo." *Psychological Bulletin* 90/2:218–44.

Coste, D. 1976. *Un niveau seuil*. Strasbourg: Council of Europe.

Council of Europe. 1991. "Language Learning for European Citizenship." Report on Workshop 1A, Curriculum Development for Modern Languages in Upper Secondary General, Technical, and Vocational Education, from Age 15/16 to Age 18/19, Rolduc, Kerkrade (Netherlands), October 21–26, 1990. *Council of Europe CC-LANG* (91).

———. 1992. *Transparency and Coherence in Language Learning in Europe: Objectives, Assessment, and Certification*. Proceedings of the Intergovernmental Symposium at Rüschlikon, Switzerland, November 10–16, 1991. Strasbourg: Council of Europe.

Criper, C., and A. Davies. 1988. *ELTS Validation Project Report*. London: British Council and Cambridge Local Examinations Syndicate.

Cronbach, L. J. 1961. *Essentials of Psychological Testing*. 2d ed. London: Harper and Row. Cited in Davies 1988.

Cummins, J. 1979. "Cognitive Academic Language Proficiency, Linguistic Interdependence, the Optimum Age Question, and Some Other Matters." *Working Papers on Bilingualism* 19:197–205.

———. 1980. "The Cross-lingual Dimensions of Language Proficiency: Implications for Bilingual Education and the Optimal Age Issue." *TESOL Quarterly* 14:175–87.

———. 1983. "Language Proficiency and Academic Achievement." In Oller 1983, pp. 108–30.

Dandonoli, P. 1987. "ACTFL's Current Research in Proficiency Testing." In Byrnes and Canale 1987, pp. 75–98.

Dandonoli, P., and G. Henning. 1990. "An Investigation of the Construct Validity of the ACTFL Proficiency Guidelines and Oral Interview Procedure." *Foreign Language Annals* 23/1:11–22.

Davidson, F., and G. Henning. 1985. "A Self-Rating Scale of English Proficiency: Rasch Scale Analysis of Items and Rating Categories." *Language Testing* 2/2.

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

——. 1991. "Language Testing in the 1990s." In Alderson and North 1991, pp. 136–52.

De Jong, J. H. A. L. 1986. "Achievement Tests and National Standards." *Studies in Educational Evaluation* 12/3:205–304.

DLI (Defense Language Institute). 1991. Interagency Language Roundtable Language Skill Level Descriptions and Appendix A from DLIFLC Pamphlet 350-14, March 1991.

Douglas, D., and L. Selinker. 1985. "Principles for Language Tests within the 'Discourse Domains' Theory of Interlanguage: Research, Test Construction, and Interpretation." *Language Testing* 2:203–26.

Dubacher, R. 1989. "L'examen oral nouvelle formule. Considérations et matériel concernant la deuxième partie de l'examen: 'Conversation.'" Ecole professionelle commerciale (EPC) Moutier.

Edgeworth, F. Y. 1890. "The Element of Chance in Competitive Examinations." *Journal of the Royal Statistical Society* 53:460–75. Cited in Linacre 1989, pp. 10–11.

Einhorn, H. L. 1974. "Expert Judgement: Some Necessary Conditions and an Example." *Journal of Applied Psychology* 59/5:562–71.

Eisenstein, M., and R. Starbuck. 1989. "The Effect of Emotional Investment on L2 Production." In *Variation in Second Language Acquisition, Volume 2: Psycholinguistic Issues*, ed. S. Gass, C. Madden, C. Preston, and L. Selinker. Clevedon, Avon (England): Multilingual Matters. Cited in Brindley 1991.

Ellis, R. 1986. *Understanding Second Language Acquisition*. Oxford: Oxford University Press.

——. 1987. "Interlanguage Variability in Narrative Discourse: Style Shifting in the Use of the Past Ter se." *Studies in Second Language Acquisition* 9/1:1–19.

ELTDU (English Language Teaching Development Unit). 1975. Stages of attainment scale.

ELTS (English Language Testing Service). See Carroll 1978.

Elviri, F.; M. G. Longhi; F. Quartepelle; and M. C. Rizzardi. 1986/87. "La ristrutturazione per livelli linguistico-comunicativi dei corsi di lingue straniere per adulti." In *L'insegnamento delle lingue straniere agli adulti*. Milan: Franco Angeli Libri, 1986. Trans. in Van Ek 1987.

Engelhard, G. 1991. "The Measurement of Writing Ability with a Many-faceted Rasch Model." Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April.

English Speaking Union Framework. See Carroll and West 1989.

Eurocentres scales of language proficiency. See North 1991, 1992a.

Faerch, C., and G. Kasper. 1983. "Plans and Strategies in Foreign Language Communication." In *Strategies in Interlanguage Communication*, ed. C. Faerch and G. Kasper, pp. 20–60. London: Longman.

Finn, R. H. 1972. "Effects of Some Variations in Rating Scale Characteristics on the Mean and Reliabilities of Ratings." *Educational and Psychological Measurement* 32:255–65.

Finnish Foreign Language Diploma for Professional Purposes. 1992. Appendix to Sajavaara 1992.

Fisher, A. G. 1991. "Development of a Functional Assessment That Adjusts Ability Measures for Task Challenge and Rater Leniency." Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April.

Foreign Service Institute (FSI) global definitions of absolute proficiency in speaking and reading. In Wilds 1975, Appendix in Palmer, Groot, and Trosper 1981.

Fredericksen, J., and A. Collins. 1989. "A System Approach to Educational Testing." *Educational Researcher* 18/9:507–8.

Freyd, M. 1923. "The Graphic Rating Scale." *Journal of Educational Psychology* 14:83–102. Cited in Landy and Farr 1983.

Glaser, R. 1963. "Instructional Technology and the Measurement of Learning Outcomes." *American Psychologist* 18:519–21. Cited in Hambleton 1978; Glass 1978; Brindley 1991.

Glass, G. V. 1978. "Standards and Criteria." *Journal of Educational Measurement* 15:237–61.

Goldstein, H. 1981. "Limitations of the Rasch Model for Educational Assessment." In Lacey and Lawton 1981, pp. 172–88.

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

Gruenfeld, E. F. 1981. *Performance Appraisal: Promise and Peril.* Ithaca, N.Y.: Cornell University Press. Cited in Linacre 1989, p. 10.

Hambleton, R. K. 1978. "Test Score Validity and Standard Setting Methods." In *Criterion-referenced Measurement*, ed. R. A. Berk, pp. 95–115. Baltimore: Johns Hopkins Press.

——. 1988. "Criterion-referenced Measurement." In Keeves 1988, pp. 277–82.

Hamp-Lyons, L., and G. Henning. 1991. "Communicative Writing Profiles: An Investigation of the Transferability of a Multiple-Trait Scoring Instrument across ESL Writing Assessment Contexts." *Language Learning* 41/3:337–73.

Harley, B.; P. Allen; J. Cummins; and M. Swain. 1987. *The Development of Bilingual Proficiency: Final Report*. Volume 1, *The Nature of Proficiency*. Toronto: Modern Language Centre OISE. Cited in Schachter 1990.

——. 1990. *The Development of Second Language Proficiency*. Cambridge: Cambridge University Press.

Harris, J.; S. Laan; and L. Mossenson. 1988. "Applying Partial Credit Analysis to the Construction of Narrative Writing Tests." *Applied Measurement in Education* 1:335–46.

Henning, G. 1984. "Advantages of Latent Trait Measurement in Language Testing." *Language Testing* 1/2:123–33.

Henning, G.; T. Hudson; and J. Turner. 1985. "Item Response Theory and the Assumptions of Unidimensionality for Language Tests." *Language Testing* 2:141–54.

Hill, R. A. 1991. "The ToPE, Test of Proficiency in English: The Development of an Adaptive Test." In Alderson and North 1991, pp. 237–46.

Hozayin, R. 1987. "The Graphic Representation of Language Competence: Mapping EFL Proficiency Using a Multidimensional Scaling Technique." In *Language Testing Research*, pp. 39–59, selected papers from the Ninth Language Testing Research Colloquium, Monterey, California, February 27–28, 1986.

Hubbard, J. P. 1971. *Measuring Medical Education*. Philadelphia: Lea and Febiger.

IBM France performance charts. In Trim 1978.

IELTS (International English Language Testing Service). Formerly ELTS. Reviewed in *Reviews of English Language Proficiency Tests*, ed. J. C. Alderson, K. J. Krahnke, and C. W. Stansfield. Washington, D.C.: TESOL, 1987.

Ingram, D. 1990. "The International English Language Testing System (IELTS): Its Nature and Development." Paper presented at the Regional Language Seminar on Language Testing and Programme Evaluation, Singapore, April 9–13.

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

Ingram, D., and E. Wylie. 1989. "Developing Proficiency Scales for Communicative Assessment." Paper presented at the National Assessment Consultation for the National Assessment Framework for Languages at Senior Secondary Level, Sydney, Australia, December 5.

Interagency Language Roundtable (ILR). 1983. Language Skill Level Descriptions and Appendix A from DLIFLC Pamphlet 350-14, March 1991.

Ivancevich, J. M. 1979. "Longitudinal Study of the Effects of Rater Training on Psychometric Error in Ratings." *Journal of Applied Psychology* 64/5:502–8.

Jacobs, R.; D. Kafry; and S. Zedeck. 1980. "Expectations of Behaviorally Anchored Rating Scales." *Personnel Psychology* 33:595–640.

Jason, H. 1962. "A Study of Medical Teaching Practices." *Journal of Medical Education* 37:1258–84.

Jones, R. L. 1985. "Some Basic Considerations in Testing Oral Proficiency." In *New Directions in Language Testing*, ed. Y. P. Lee et al., pp. 77–84. New York: Pergamon.

Kavanagh, M. J.; A. C. MacKinney; and L. Wolins. 1971. "Issues in Managerial Performance: Multitrait-Multimethod Analysis of Ratings." *Psychological Bulletin* 75:34–49. Cited in Landy and Farr 1983.

Keaveny, T. J., and A. F. McGann. 1975. "A Comparison of Behavioral Expectation Scales." *Journal of Applied Psychology* 60:695–703.

Keeves, J. P., ed. 1988. *Educational Research, Methodology, and Measurement: An International Handbook.* New York: Pergamon.

Kenyon, D. M., and C. W. Stansfield. 1992. "Examining the Validity of a Scale Used in a Performance Assessment from Many Angles Using the Many-Faceted Rasch Model." Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Kingstrom, P. O., and A. R. Bass. 1981. "A Critical Analysis of Studies Comparing Behaviorally Anchored Rating Scales (BARS) and Other Rating Formats." *Personnel Psychology* 34:263–89.

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

Kruskal, J. B., and M. Wish. 1978. *Multidimensional Scaling.* A Sage University Paper. Beverly Hills, Calif.: Sage Publications.

Lacey, C., and D. Lawton, eds. 1981. *Issues in Evaluation and Accountability.* London: Methuen.

Lambert, R. 1992. Personal communication.

Lance, C. E.; M. S. Teachout; and T. M. Donnelly. "Specification of the Criterion Construct Space: An Application of Hierarchical Confirmatory Factor Analysis." *Journal of Applied Psychology* 77/4:437–52.

Landy, F. L., and J. L. Farr. 1980. "Performance Rating." *Psychological Bulletin* 87/1:72–102.

———. 1983. *The Measurement of Work Performance: Methods, Theory, and Applications.* San Diego: Academic Press.

Lantolf, J. P., and W. Frawley. 1984. "Speaking and Self Order: A Critique of Orthodox L2 Research." *Studies in Second Language Acquisition* 6/2:143–59.

———. 1985. "Oral Proficiency Testing: A Critical Analysis." *Modern Language Journal* 70:337–45.

———. 1988. "Proficiency: Understanding the Construct." *Studies in Second Language Acquisition* 10/2:181–96.

———. 1992. "Rejecting the OPI—Again: A Response to Hagen." *ADFL Bulletin* 23/2:34–37.

Linacre, J. M. 1988. *FACETS: A Computer Program for the Analysis of Multi-faceted Data.* Chicago: MESA Press.

———. 1989. *Multi-faceted Measurement.* Chicago: MESA Press.

———. 1992. Personal communication.

Liskin-Gasparro, J. E. 1984. "The ACTFL Proficiency Guidelines: A Historical Perspective." In *Teaching for Proficiency, the Organizing Principle,* ed. T. V. Higgs, pp. 11–43. Lincolnwood, Ill.: National Textbook Company.

Lissitz, R. W., and S. B. Green. 1975. "Effect of the Number of Scale Points on Reliability: A Monte Carlo Approach." *Journal of Applied Psychology* 60:10–13. Cited in McKelvie 1978.

Lowe, P., Jr. 1983. "The ILR Oral Interview: Origins, Applications, Pitfalls, and Implications." *Unterrichtspraxis* 16/2:230–44.

———. 1985. "The ILR Proficiency Scale as a Synthesising Research Principle: The View from the Mountain." In *Foreign Language Proficiency in the Classroom and Beyond*, ed. C. J. James, pp. 9–53. Lincolnwood, Ill.: National Textbook Company.

McKelvie, S. J. 1978. "Graphic Rating Scales—How Many Categories?" *British Journal of Psychology* 69:185–202.

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

Mareschal, R. 1977. "Normes linguistiques: Détermination, description, contenu, utilité." *Canadian Modern Language Review* 33:620–31.

Marsh, H. W., and D. Hocevar. 1988. "A New, More Powerful Approach to Multitrait Multimethod Analyses: Application of Second Order Confirmatory Factor Analysis." *Journal of Applied Psychology* 73:107–17. Cited in Lance, Teachout, and Donnelly 1992.

Masters, G. 1988. "Partial Credit Model." In Keeves 1988, pp. 292–97.

Matell, M. S., and J. Jacoby. 1971. "Is There an Optimal Number of Categories for Likert Scale Items? Study 1: Reliability and Validity." *Educational Psychological Measurement* 31:657–74. Cited in McKelvie 1978.

Meisel, J.; H. Clahens; and M. Pienemann. 1981. "On Determining Developmental Stages in Second Language Acquisition." *Studies in Second Language Acquisition* 3/2:109–35.

Meredith, R. A. 1990. "The Oral Proficiency Interview in Real Life: Sharpening the Scale." *Modern Language Journal* 74/3:288–96.

Messick, S. 1989. "Meaning and Values in Test Validation: The Science and Ethics of Assessment." *Educational Researcher* 18/2:5–11.

Milanovic, M.; N. Saville; A. Pollitt; and A. Cook. 1992. "Developing and Validating Rating Scales for CASE: Theoretical Concerns and Analyses." Paper presented at the Fourteenth Language Testing Research Colloquium, Vancouver, February 27–March 1.

Miller, G. A. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *Psychological Review* 63:81–97. Cited in Finn 1972.

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

———. 1986. "The Evaluation of Tests of Communicative Performance." In *Innovations in Language Testing*, ed. M. Portal, pp. 1–13. Windsor, England: NFER-Nelson.

Munby, J. 1978. *Communicative Syllabus Design*. Cambridge: Cambridge University Press.

Murphy, K. R., and R. L. Anhalt. 1992. "Is Halo Error a Property of Rater, Ratees, or the Specific Behaviors Observed?" *Journal of Applied Psychology* 77/4:494–500.

Murphy, K. R., and J. I. Constans. 1987. "Behavioral Anchors as a Source of Bias in Rating." *Journal of Applied Psychology* 72/4:573–77.

Murphy, K. R., and V. A. Pardaffy. 1989. "Bias in Behaviorally Anchored Rating Scales: Global or Scale Specific." *Journal of Applied Psychology* 74/2:343–46.

Muzzin, L. J., and E. Hart. 1985. "Oral Examinations." In *Assessing Clinical Competence*, ed. V. R. Neufeld and G. R. Normal. New York: Springer.

Myford, C. M. 1991. "Assessment of Acting Ability." Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, April.

National Curriculum Council. 1991. *National Curriculum Council Consultation Report: Modern Foreign Languages in the National Curriculum*. York, England: National Curriculum Council.

Norman, W. T. 1963. "Towards an Adequate Taxonomy of Personality Attributes: Replicated Factor Structures in Peer Nomination Personality Ratings." *Journal of Abnormal and Social Psychology* 66:574–83. Cited in Cooper 1981.

Norman, W. T, and L. R. Goldberg. 1966. "Rater, Ratees, and Randomness in Personality Structure." *Journal of Personality and Social Psychology* 4:681–91.

North, B. 1986. "Activities for Continuous Communicative Assessment." Unpublished M.A. Phase 3 project, English Language Research Dept., University of Birmingham.

——. 1991. "Standardisation of Continuous Assessment Grades." In Alderson and North 1991, pp. 167–77.

——. 1992a. "Activités de communication et évaluation de l'oral." *Autour de l'évaluation: Bulletin CILA* (Neuchâtel) 55:97–108.

——. 1992b. "A European Language Portfolio: Options for Scales for Proficiency." Paper presented at the Intergovernmental Symposium "Transparency and Coherence in Language Learning in Europe: Objectives, Assessment, and Certification," Rüschlikon, Switzerland, November 10–16, 1991. In Council of Europe 1992. Rpt. in Schärer and North 1992.

——. 1992c. "Kompetenzbeschreibung: Fertigkeitsskalen und Bewertungskriterien für den Fremdsprachenunterricht." Paper presented at the Symposium to Launch a Swiss Framework Project, Fribourg, Switzerland, November.

North, B.; B. Page; L. Porcher; G. Schneider; and J. Van Ek. 1992. "A Preliminary Investigation of the Possible Outline Structure of a Common European Language Framework: Synthesis of the Outcome of the Short Intensive Meeting of Invited Experts to Consider Principles, Hypotheses, and Options for Further Investigation." *Council of Europe CC-LANG* (92) 12.

Oller, J. 1976. "Evidence for a General Language Proficiency Factor: An Expectancy Grammar." *Die Neueren Sprachen* 2:165–74. Rpt. in Oller 1983, pp. 3–10.

——, ed. 1983. *Issues in Language Testing Research*. Rowley, Mass.: Newbury House.

Page, B., and D. Hewett. 1987. *Languages Step by Step: Graded Objectives in the UK*. London: Centre for Information on Language Teaching and Research.

Palmer, A. S.; J. M. Groot; and G. A. Trosper. 1981. *The Construct Validation of Tests of Communicative Competence*. Washington, D.C.: TESOL.

Passini, F. T., and W. T. Norman. 1966. "A Universal Conception of Personality Structure?" *Journal of Personality and Social Psychology* 4:44–49. Cited in Cooper 1981.

Paterson, D. G. 1922. "The Scott Company Graphic Rating Scale." *Journal of Personnel Research*. Cited in Landy and Farr 1980.

Paulston, C. B. 1990. "Educational Language Policies in Utopia." In Harley et al. 1990, pp. 187–200.

Pienemann, M., and M. Johnson. 1987. "Factors Influencing the Development of Language Proficiency." In *Applying Second Language Acquisition Research*, ed. D. Nunan, pp. 45–141. Adelaide: National Curriculum Resource Centre.

Pienemann, M.; M. Johnson; and G. Brindley. 1988. "Constructing an Acquisition-based Procedure for Second Language Assessment." *Studies in Second Language Acquisition* 10/2: 217–44.

Pollitt, A. 1991. "Reply to Alderson, 'Bands and Scores.'" In Alderson and North 1991, pp. 87–94.

———. 1992. Personal communication.

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

Popham, W. J. 1978. *Criterion-referenced Measurement*. Englewood Cliffs, N.J.: Prentice-Hall. Cited in Hambleton 1978.

Porter, D. 1991. "Affective Factors in Language Testing." In Alderson and North 1991, pp. 32–40.

Raffaldini, T. 1988. "The Use of Situation Tests as Measures of Communicative Ability." *Studies in Second Language Acquisition* 10/2:197–216.

Rampton, B. 1987 "Stylistic Variability and Not Speaking Normal English: Some Post-Labovian Approaches and Their Implications for the Study of Interlanguage." In *Second Language Acquisition in Context*, ed. R. Ellis, pp. 47–58. Englewood Cliffs, N.J.: Prentice-Hall. Cited in Brindley 1991.

Raymond, M. R.; L. C. Webb; and W. C. Houston. 1991. "Correcting Performance Rating Errors in Oral Examinations." *Evaluation and the Health Professions* 14:100–122.

Royal Society of Arts (RSA). 1989. Examinations Board, Modern Languages Examinations Explanatory Booklet and Syllabus Guidelines: Characteristics of the Performance Expected from Candidates at the Various Levels.

Ruggles, A. M. 1911. *Grades and Grading*. New York Teachers' College. Cited in F. J. Kelly, *Teacher's Marks*. New York Teacher's College, 1914. Cited in Linacre 1989, p. 10.

Saal, F. E.; R. G. Downey; and M. A. Lahey. 1980. "Rating the Ratings: Assessing the Psychometric Quality of Rating Data." *Psychological Bulletin* 88/2:413–28.

Sajavaara, K. 1992. "Designing Tests to Meet the Needs of the Workplace." In Shohamy and Walton 1992.

Savignon, S. J. 1992. "This Is Only a Test: What Classroom Tests Tell Learners about Language and Language Learning." In Shohamy and Walton 1992.

Schachter, J. 1990. "Communicative Competence Revisited." In Harley et al. 1990, pp. 50–56.

Schärer, R. 1992. "A European Language Portfolio." In Council of Europe 1992. Rpt. in Schärer and North 1992.

Schärer, R., and B. North. 1992. *Towards a Common European Framework for Reporting Language Competency.* NFLC Occasional Paper. Washington, D.C.: National Foreign Language Center.

Schneider, G., and R. Richterich. 1992. "Transparency and Coherence: Why and for Whom." Paper presented at the Intergovernmental Symposium "Transparency and Coherence in Language Learning in Europe: Objectives, Assessment, and Certification," Rüschlikon, Switzerland, November 10–16, 1991. In Council of Europe 1992.

SCOTVEC (Scottish Vocational Educational Council). 1989. Higher National Unit Specification; National Certificate Module Descriptor.

Shohamy, E. 1988. "A Proposed Framework for Testing the Oral Language of Second/Foreign Language Learners." *Studies in Second Language Acquisition* 10/2:165–79.

Shohamy, E., and R. Walton, eds. 1992. *Language Assessment for Feedback: Testing and Other Strategies.* Washington, D.C.: National Foreign Language Center/Kendall Hunt.

Sinclair, J. McH. 1979. "Applied Discourse Analysis—An Introduction" and "Some Implications of Discourse Analysis for ESP Methodology." *Applied Linguistics* 1/3:185–88, 253–61.

——. 1981. "Planes of Discourse." Unpublished paper.

——. 1985. Personal communication.

Sinclair, J. McH., and M. Coulthard. 1975. *Towards an Analysis of Discourse.* London: Oxford University Press.

——. 1982. *Teacher Talk.* Oxford: Oxford University Press.

Skaggs, G., and R. W. Lissitz. 1986. "IRT Test Equating: Relevant Issues and a Review of Recent Research." *Review of Educational Research* 56/4:495–529.

——. 1988. "Effect of Examinee Ability on Test Equating Invariance." *Applied Psychological Measurement* 12/1:69–82.

Skehan, P. 1984. "Issues in the Testing of English for Specific Purposes." *Language Testing* 1/2:202–20.

———. 1988. "Language Testing: Survey Article Part 1." *Language Teaching* 21/4:211–21.

Smith, P. C., and J. M. Kendall. 1963. "Retranslation of Expectations: An Approach to the Construction of Unambiguous Anchors for Rating Scales." *Journal of Applied Psychology* 47/2:149–55.

Spolsky, B. 1992. "Testing and Examinations in a National Foreign Language Policy." Paper presented at the International Conference on Foreign Language Policies, sponsored by the National Foreign Language Center and the University of Jyväskalä, Jyväskalä, Finland, August.

Stahl, J. A., and M. Lunz. 1991. "Judge Performance Reports: Medium and Message." Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Stahl, J. A.; M. Lunz; and B. D. Wright. 1991. "Equating Examinations That Include Judges (Multiple Facets)." Paper presented at the National Council of Measurement in Education, Chicago, April.

Stern, H. H. 1983. "Models of Second Language Learning and the Concept of Proficiency." Chapter 16 of *Fundamental Concepts of Language Teaching*. Oxford: Oxford University Press.

Tall, G. 1981. "The Possible Dangers of Applying the Rasch Model to School Examinations and Standardised Tests." In Lacey and Lawton 1981, pp. 189–203.

Tarone, E. 1983. "On the Variability in Interlanguage Systems." *Applied Linguistics* 4:142–63.

Theunissen, T. J. J. M. 1987. "Text Banking and Test Design." *Language Testing* 4/1:1–8.

Thurstone, L. L. 1928a. "Attitudes Can Be Measured." *American Journal of Sociology* 33:529–54. Cited in Wright and Masters 1982, p. 15.

——. 1928b. "The Measurement of Opinion." *Journal of Abnormal and Social Psychology* 22:415–30. Cited in Wright and Masters 1982, p. 5.

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

——. 1991. Personal communication.

——. 1992. "Report on Two Preliminary Meetings on the Follow-up to Be Given to the Rüschlikon Symposium." *Council of Europe CC-LANG* (92) 3.

Tyndall, B., and D. M. Kenyon. Forthcoming. "Validation of a New Holistic Rating Scale Using Rasch Multi-faceted Analysis." Unpublished paper.

University of Cambridge/RSA (Royal Society of Arts). 1990. Certificates in Communicative Skills in English, Teachers' Guide: "Degrees of Skill" for Levels 1, 2, 3, 4 for Listening, Reading, Speaking, and Writing.

Van Ek, J. A. 1975. *The Threshold Level*. Strasbourg: Council of Europe.

——. 1986. *Objectives for Foreign Language Teaching, Volume 1: Scope*. Strasbourg: Council of Europe.

——. 1987. *Objectives for Foreign Language Teaching, Volume 2: Levels*. Strasbourg: Council of Europe.

Van Ek, J. A., and J. L. M. Trim. 1990. *The Threshold Level 1990*. Strasbourg: Council of Europe.

Van Lier, L. 1989. "Reeling, Writhing, Fainting, and Stretching in Coils: Oral Proficiency Interviews as Conversation." *TESOL Quarterly* 23:489–508.

Vonarburg, B. 1992. "Points of Encounter." In Council of Europe 1992.

Walther, R. 1991. *Beurteilung von Schülerleistungen im Fach Französisch am Ende der Sekundarstufe I.* Bern: Amt für Bildungsforschung des Kantons Bern.

Walton, R. 1992. Personal communication.

Warmke, D. L., and R. S. Billings. 1979. "Comparison of Training Methods for Improving the Psychometric Quality of Experimental and Administrative Performance Ratings." *Journal of Applied Psychology* 64:124–31. Cited in Cooper 1981.

Weir, C. 1989. *Communicative Language Testing.* Exeter: University of Exeter Press.

Westaway, G. 1988. "Developments in the English Language Testing Service (ELTS) M2 Writing Test." *Australian Review of Applied Linguistics* 11/2:13–29.

Westaway, G.; J. C. Alderson; and C. M. Clapham. 1990. "Directions in Testing for Specific Purposes." In *Individualising the Assessment of Language Abilities,* ed. P. De Jong and D. P. Stevenson, pp. 239–56. Clevedon, Avon (England): Multilingual Matters.

Wherry, R. J. 1952. *The Control of Bias in Rating: A Theory of Rating.* Personnel Board Report 922. Washington, D.C.: Department of the Army, Personnel Research Section. Cited in Landy and Farr 1980, 1983.

Widdowson, H. G. 1978. *Teaching Language as Communication.* Oxford: Oxford University Press.

——. 1979. *Explorations in Applied Linguistics.* Oxford: Oxford University Press.

Wilds, C. P. 1975. "The Oral Interview Test." In *Testing Language Proficiency,* ed. B. Spolsky and R. Jones. Washington D.C.: Center for Applied Linguistics.

Wilson, M. 1989. "Empirical Examination of a Learning Hierarchy Using an Item Response Theory Model." *Journal of Experimental Education* 57/4:357–71.

Wolf, R. M. 1988. "Rating Scales." In Keeves 1988, pp. 496–97.

Woods, A., and R. Baker. 1985. "Item Response Theory." *Language Testing* 2:117–40.

Wright, B. D., and G. Masters. 1982. *Rating Scale Analysis: Rasch Measurement.* Chicago: MESA Press.

Yorozuya, R., and J. W. Oller. 1980. "Oral Proficiency Scales: Construct Validity and the Halo Effect." *Language Learning* 30:135–53.

Zuengler, J. 1989. "Performance Variation in NS-NNS Interactions: Ethnolinguistic Difference or Discourse Domain?" In *Variation in Second Language Acquisition, Volume 1: Discourse and Pragmatics,* ed. S. Gass, C. Madden, C. Preston, and L. Selinker, pp. 228–43. Clevedon, Avon (England): Multilingual Matters. Cited in Brindley 1991.

The National Foreign Language Center

at the Johns Hopkins University
1619 Massachusetts Avenue NW
Washington DC 20036