DOCUMENT RESUME

ED 380 983 FL 021 922

AUTHOR Hart-Gonzalez, Lucinda

TITLE Raters and Scales in Oral Proficiency Testing: The FSI Experience.

PUB DATE 5 Mar 94
NOTE 26p.; Paper presented at the Annual Language Testing Research Colloquium (Washington, DC, March 5, 1994).
PUB TYPE Speeches/Conference Papers (150) -- Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Federal Government; Interrater Reliability; *Language Proficiency; *Language Tests; *Oral Language; *Rating Scales; Scores; Second Language Learning; *Second Languages; Test Construction; Testing
IDENTIFIERS *Foreign Service Institute VA; *Oral Proficiency Testing

ABSTRACT

This overview of the 40 year development of language proficiency rating at the Foreign Service Institute (FSI) and elsewhere in the federal government focuses on three issues pertaining to the scale and the raters: (1) the number of levels of differentiation in thc scale; (2) the relation of the scale to the rating task; and (3) the calibration of the scale against some underlying proficiency continuum. Over the years, the rating system has become more and more complex, all in a seeming effort to balance between the desire for greater differentiation and information, concern for rater fairness, and the administrative need for a simple but dependable global score. Problems occur when greater specificity contradicts scale assumptions or imposes undesirable ones. At this writing, The Federal Language Testing Board, which includes FSI, is revising the testing system in the direction of greater simplicity for raters, with some attention to issues of scale. Contains 26 references. (Author)

RATERS AND SCALES IN ORAL PROFICIENCY TESTING: THE FSI EXPERIENCE

Lucinda Hart-Gonzalez
Foreign Service Institute
National Foreign Affairs Training Center
4000 Arlington Blvd.
Arlington, VA 22204-1500

Presented at:
Language Testing Research Colloquium
Center for Applied Linguistics
Washington, DC
March 5, 1994

Abstract: This overview of the forty year development of language proficiency rating at the Foreign Service Institute and elsewhere in the federal government focuses on three issues pertaining to the scale and the raters: (1) the number of levels of differentiation in the scale, (2) the relation of the scale to the rating task, and (3) the calibration of the scale against some underlying proficiency continuum. Over the years the rating system has become more and more complex, all in a seeming effort to balance between the desire for greater differentiation and information, concern for rater fairness, and the administrative need for a simple but dependable global score. Problems occur when greater specificity contradicts scale assumptions or imposes undesirable ones. At this writing, the Federal Language Testing Board, which includes FSI, is revising the testing system in the direction of greater simplicity for raters, with some attention to issues of scale.

RATERS AND SCALES IN ORAL PROFICIENCY TESTING: THE FSI EXPERIENCE

Lucinda Hart-Gonzalez

RATERS AND SCALES IN ORAL PROFICIENCY TESTING: THE FSI EXPERIENCE [1]

Lucinda Hart-Gonzalez
Foreign Service Institute
National Foreign Affairs Training Center
4000 Arlington Blvd.
Arlington, VA 22204-1500

1. INTRODUCTION

1.1. Global vs. Detailed Ratings

Within the federal context, the importance of proficiency test scores is accentuated for several reasons. First of all, language training in the government is not for hypothetical use in the distant future; it is of national importance that we be able to gauge whether the language proficiency of our civilian and military personnel is adequate to their missions. Second, from the point of view of the employees themselves, language proficiency ratings may play a role in job selection, promotion and job security; from this perspective, issues of scale such as "How close is close enough" become critical to thousands in the military, foreign and civil service.

Language proficiency test scores are also the essential "bottom line" in nearly all federal research on language training. In basic institutional and applied research, such as progress reports on students in training, program reviews of language section performance, or tests of the effectiveness of new materials or techniques, language proficiency is the metric on which success is measured. In advanced research, such as efforts to predict language learning rate from student background variables, language proficiency is again a key metric.

The Interagency Language Roundtable (ILR) scale of language proficiency was originally developed at the Foreign Service Institute and is still known in some circles as the FSI scale. It is also the basis for the ACTFL scale (American Council on the Teaching of Foreign Languages). The ILR scale is composed of six basic levels:

0	'No proficiency'
1	'Limited Proficiency'
2	'Limited Working Proficiency'
3	'Full Working Proficiency'
4	'Advanced Working Proficiency'
5	'Educated Native Speaker'

Additionally, 'plus'-levels on 0-4, e.g. 1+, indicate that some, but not all, aspects of the next higher level are present. There are ILR criteria for Speaking, Reading, for non-interactive Listening, and for Writing. Criteria are currently being developed for Translation (Cascallar et al. 1993). FSI does not test for non-interactive Listening, but includes interactive listening in its Speaking scale. The ACTFL scale of 'Novice' through 'Superior' adds further differentiation to the ILR 0-3.

What follows is a review of the development and use of the ILR scale at the Foreign Service Institute and other Roundtable agencies in terms of three psychometric issues: the number of levels of differentiation in the scale, the relation of the scale to the rating process, and the calibration of scale.

The first issue concerns **the number of levels of differentiation in the scale**. Klein-Braley (1991:81) recommends a wide range of scores. Greater differentiation enables statistical analysis to highlight patterns which might otherwise be missed. Likewise, it prevents claims that are too strong, where patterns or relationships may not be as fine-grained as first thought. For administrative purposes, on the other hand, virtue is on the side of a simpler scale with fewer levels for easier decisions.

The primary purpose of greater differentiation in the federal context is rater training and support. Additional steps in the rating process are thought to help guide the raters' attention to all aspects of the test sample, to help confirm or reject tentative ratings, and to support final ratings.

While "fine-tuning" may be seen as an aide, reliability falls when the scale itself becomes too fine-grained. At FSI, the six-point 0-5 scale (or perhaps 11-point with the plus-levels) is applied using, as a guide, a series of further refining six-point scales, which are then further scaled, ending up with Index Scores ranging from 0-120. As we follow the evolution of language testing at FSI, we shall see this development as part of an ongoing tension between the poles of generalization and differentiation.

A second issue involves **the relation of the scale to the rating task.** One way for the rater to apply the scale is holistically; another is to apply it first to a set of component criteria from which the global rating is derived. Still a third is to use the same scale for the various components, but not to assign a single global rating at all, leaving a profile of ratings. The second or third systems, involving component ratings, are another way of affording greater differentiation, but without necessarily increasing the number of levels in the scale. Greater differentiation should mean more information. On the other hand, some studies suggest that both inter- and intra-rater reliability drop for component ratings in comparison to holistic ratings (Littlefield & Troendle 1987, 1986).

Bachman (1993, 1988) advocates the third approach, a language proficiency profile in which various facets of proficiency are rated separately. The advantages of such a system are both theoretical and practical. In terms of theory, by separating the scales, we avoid the assumption that all skills progress in the same way, or even in the same direction. For instance, fluency may drop when tackling a highly spontaneous and informal speaking style, even though the style has a limited (though idiomatic) vocabulary.

One practical advantage of having a profile of ratings is that it assists teachers in evaluating learner progress and diagnosing possible problems. Adams (1980) had found that the factors which most discriminated between different levels of end-of-training proficiency differed from level to level. For instance, vocabulary was the single most important factor distinguishing the 0+ from the 1-level speaker. Fluency distinguished the 1 from the 1+ and comprehension the 1+ from the 2. Higgs and Clifford (1982) suggested on the basis of ILR component scores that so-called "terminal 2's," those who reach the 2 or 2+ but never advance to the 3, are marked by weakness in grammatical structure, whatever their strengths in vocabulary and discourse strategies. Grammar was also the leading distinguishing factor between the 2+ and the 3 in Adams' study.

At their most extreme, the federal agencies follow the first approach, the global-only. Agencies need a global rating for the various kinds of decisions noted above. Not only do factor scores have no administrative meaning, even Reading and Speaking are generally interpreted as a unit. For example, most language-designated positions (LDPs) in the Foreign Service require a 3/3 (i.e. S-3/R-3 for Speaking/Reading). The 3/3 is a threshold; anything less is generally held insufficient to leave for an overseas post.

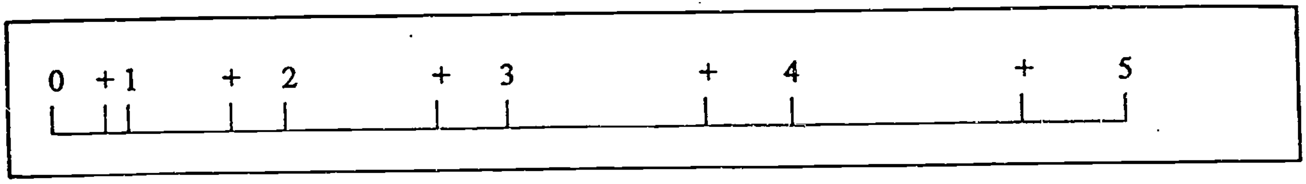
In practice, the federal agency testers follow Approach Two. All have a system of component (e.g. structure, vocabulary, fluency) and global ratings for each of those skills. The interplay is iterative, a back-and-forth between global rating and factor ratings. This is consistent with the finding that global ratings as initial hypotheses are a natural rater process (Littlefield and Troendle 1986) and are more consistently achieved, even after rater training, than are compononent ratings, especially in the middle ranges of the scale (Braungart-Bloom 1986).

Beyond these similarities, the agencies do handle the global and factor ratings somewhat differently. In FSI testing, the factor ratings interact with the global ratings via 120 levels of the Index score. The partial scores on Performance Levels in the factors are more cognitively manageable, a move toward the simplicity end of the tension, while still preserving differentiation.

A third question, in many ways dependent on the first two, concerns **the calibration of the scale**. Evidence suggests that the ILR scale is at least ordinal (Clark & Lett 1988:78); that is, a 2 is higher than a 1, but not by a specific amount. Thus, if one examinee got to a 2 and another got to a 1, both starting from 0, we cannot say that the first did twice as well as the second. This is the most conservative interpretation of the scale.

There is further agreement that the intervals at which these proficiency scores map onto the underlying theoretical scale are larger at the higher levels than at the lower levels. That means that while the numbers used to represent scores increase in equal intervals, the distances they represent on the underlying scale are actually expanding. The Defense Language Institute (DLI) maps this scale as shown in Figure 1. I have proposed a useful formula for transforming basic rating scores to reflect that expanding scale, namely (Score + 1)Squared (Hart-Gonzalez, 1993a,b).

Figure 1. Defense Language Institute Representation of ILR Scale (Used in Rater Training)



The expanding-interval nature of the rating scale is affected by varying the rating process, and by the levels of differentiation used in that process. For example, while the proposed formula neatly maps the basic 0-5 levels of Figure 1, it does not handle the plus-levels as easily (a work-around is to value the plus-level as an additional .7 and then the transformation becomes: ((Score + 1) * 10)Squared but then the distances between step become highly exaggerated) (Hart-Gonzalez 1993a).

The sections of this study each highlight components introduced in the rating process during its evolution. Each section first describes the level of differentiation at that point, then examines how the raters' task of scoring relates to the scale, and then discusses the ramifications of that stage of development on calibration.

Overall, we shall see that current FSI practice is the product of decades of balanced tension between the need for diffentiation and greater information on the one hand, and for generalization and rater reliability on the other. The resulting system affords some of the best of both, but not without compromising effects on the nature of the scale.

1.2. The Historical Record

R.A.C. Goodison begins his 1946-1989 compendium of New York Times articles on FSI language training with the sad and all-too-common observation: "In government offices, institutional memories are often short or non-existent" (1990:2). Federal workers typically feel more compelled to write internal memos, manuals and reports than to publish in the scholarly press. Such internal papers are the ones most likely to disappear with personnel turnover. This account is pieced together from a few published histories and from oral accounts of some of those involved.

The written record is fairly clear on the 1950's when Frank Rice and Claudia P. Wilds first devised rating scale and oral proficiency test, based on the then revolutionary concept of the structured interview. This is followed by a relative lull in historical record during the 1960's. The 1970's, when Marianne Lehr Adams took over from Wilds as head of the FSI testing unit, saw increased academic publishing (e.g. Jones and Spolsky 1975, Clark 1978, Frith 1980) surely in part spurred by close relationships between the Interagency Language

Roundtable and Georgetown University with its annual Roundtable, the Center for Applied Linguistics, also in Washington, and the Educational Testing Service, which eventually took part in developing the ACTFL/ETS proficiency guidelines widely used in schools and colleges.

There is still much left to disentangle in the development of this history. The focus here, however, is not on chronology, but rather, on the nature of the measurement scale and the process of rating using it.

2. THE FACTORS

2.1. Differentiation

Early refinement of the scale was all toward greater differentiation. The original FSI scale was a reflection of the categories used in a government-wide survey of language ability, conducted at the start of the Cold War. The first step in the development of language proficiency scales was to find criteria general enough to be valid across all languages. Unlike in the original survey, speaking and reading were distinguished, with heavy emphasis on speaking, including listening (Claudia P. Wilds, personal communication, 2/14/94).

After describing basic levels, the Roundtable tackled the problem of border scores, and thus developed the plus-scores to differentiate those who perform at the high end of a performance level, but who cannot perform consistently at the next higher level (John L.D. Clark, personal communication, February 16, 1994). While in one sense the plus-levels are part of the basic levels and not separate level themselves, in another sense, they behave rather like five added levels of the scale. Indeed in the 1980 study cited above, Marianne Adams of the FSI Testing Unit describes the scale as "11 possible global speaking (S-) scores" (Adams 1980:1). In some agencies, administrative decisions, such as incentive pay, are based on a plus-level rating (Olga Navarrete, personal communication, February 16, 1994).

Official test scores have gone no further than this in making distinctions. Any further differentiation has been to help with rating and explaining ratings to the examinee, but is not an official score.

By 1959, Frank Rice and Claudia Wilds had finished developing and had insituted a test evaluation checklist with five factors: accent, grammar, vocabulary, fluency, and comprehension (Cornwell & Budzinski 1993ms:29). The purpose of this checklist was to ensure that ratings did not overemphasize one aspect of proficiency (say, for instance, use of the pluperfective) without due attention to all of the criteria. In the beginning, these factors were rated on a six-point scale. The scale was further demarcated using colons into either a three-point scale, as shown in the first three factors in Table I from Adams (1980:1), or a four-point scale, as shown in the last two factors in Table I from Wilds (1975:38).

Table I. The Early FSI Factor Scale

Accent	foreign	__ __ : __ __ : __ __	native
Comprehension	incomplete	__ __ : __ __ : __ __	complete
Fluency	uneven	__ __ : __ __ : __ __	even
Grammar	inaccurate	__ :__ __ : __ __ : __	accurate
Vocabulary	inadequate	__ :__ __ : __ __ : __	adequate

(Taken from Adams 1980:1, Wilds 1975:38)

With some modification, that factor checklist became the basis for the Speaking Performance Profiles used at other Roundtable agencies such as CIA, DLI and FBI (Herzog 1988:165). [2] Since those agencies test listening as a separate skill, one change they made is that comprehension is not evaluated as a component factor of the speaking test. However, those agencies added a factor for sociolinguistic or cultural proficiency. The Speaking Performance profile also lists a sixth factor, Tasks. That addition responds in some ways to Bachman's call for closer examination of performance tasks and language competencies as they relate in language testing (Bachman 1985). More bropadly, it responds to aa general shift in focus in linguistics from behavioral, structural models to cognitive, sociolinguistic models of language and language use.

The FSI factor list was also revised, in 1983-1984, as part of a larger reworking of the test system. Strongly influenced by the work of Leonard Talmy (e.g. 1978), Argoff sought to make the test more reflective of the central role of meaning in communication, and of our increased understanding of how meaning is organized in the mind. (H. David Argoff, Personal Communication, February 22, 1994). The Lexicalization factor which replaced Vocabulary is a key reflection of that change, as was the addition of Discourse, a level which includes the speaking task or text and sociocultural aspects, in addition to other above-the-sentence structure. Other factors, comprehension and fluency, were retained, while Structure expanded the traditional scope of grammar to include pronunciation. Table II lists the factors currently used by each of the agencies.

Table II Speaking Test Performance Factors

FSI	CIA/DLI/FBI
Comprehension	
Discourse	Sociolinguistic/ Cultural
	Tasks
Structure	Grammar
	Pronunciation
Lexicalization	Vocabulary
Fluency	Fluency/ Integrative

2.2. Raters

We now examine the manner in which the rating scales relate to scoring. At CIA/DLI/FBI, the global proficiency rating is assigned first. Raters then fill out the Speaking Performance Profile by marking each of the Table II factors on a line that is calibrated at equal intervals with the 0-5 scale. In essence, the rater uses the Profile to substantiate the global rating, a preliminary hypothesis, by running a checklist of test performance in greater detail. In training, raters-to-be fill out the Speaking Performance Profile first, using it as the basis for building an inner concept of the global score.

The Profile is filled out by marking along a line on each factor. The marks of the two raters are then measured by centimeter and averaged, on the assumption that "statistically the average rating is always more accurate than the rating of the best scorer" (Quinones, in Wilds 1975).

To confirm a global rating, the factor scores must generally be at least as high as the overall proficiency rating. This is particularly true of the Tasks factor, which must be the same as the overall score (Lowe, personal communication, 9/13/93). However, if just one of the other factors is slightly below the global rating, strength in another factor may compensate and the rating will hold (Gardner, personal communication, 9/20/93). At DLI, likewise, in order to assign a global rating of S-3 (S for Speaking), the rater must have assigned at least a 3 in each of the performance factors, and equivalent is true for any higher rating. For a rating of 2+ or lower, compensatory rating is allowed; that is, a high rating on one factor, say Pronunciation, may compensate for a lower rating on another factor, say Grammar (Personal communication, Pat Dege, 9/14/93).

2.3. Calibration

This system presents three scale calibration problems. The first is that the scale representations on which raters are marking the factors have equal intervals, while the scale these lines represent has expanding intervals. This leads to over-differentiation at the lower end where the line is too long, and under-differentiation at the higher end, where the line is too short. Over-differentiation at the lower end, is actually a need broadly felt in the language testing community. It motivated the ACTFL scale, and even in the federal language training community, most students are at the lower levels or would not be in language training.

The problem this miscalibration presents for the practice of averaging ratings is different. The distortion of expanding intervals into equal intervals also distorts the mean. At higher levels there will be a tendency to underrate and at lower levels, a possible tendency to overrate.

There is also a potentially more serious problem interpreting the meaning of any given score on the scale. The fact that the lower level ratings may represent an average score or a minimum score, while the higher level ratings represent a minimum score only, presents problems for the scale calibration. Test scores are meant to represent bands on some underlying continuous scale of language proficiency. When a single score is used to represent a continuous band, that single score must be representative in a systematic way, such as a mid-point or upper or lower threshold; for example, the 10-year age cohorts (10-19, 20-29, 30-39, etc) may be represented by (15, 25, 35) or by (10, 20, 30) or by (19, 29, 39), but not by (15, 20, 30). Nonetheless, this last is precisely what we get by compensating at the low end and not the high end of the scale.

3. THE INDEX SCORE

3.1. Differentiation

FSI has taken a somewhat different approach to differentiation and the global/factor relationship. By the 1960's, under Claudia Wilds, FSI developed an additive scoring system. A multiple correlation study (apparently unpublished and no longer available) was done to determine the relative contribution of each of the factors (Wilds 1975 or 1979).[3] No mention survives of whether the study also evaluated the relative independence (or lack) among the factors. In any event, weights were assigned to each of the factors as shown in Table III.

Table III Speaking Factor Weights (ca. 1967)

	1	2	3	4	5	6	
Accent	0	1	2	2	3	4	_____
Grammar	6	12	18	24	30	36	_____
Vocabulary	4	8	12	16	20	24	_____
Fluency	2	4	6	8	10	12	_____
Comprehension	4	8	12	15	19	23	_____
						Total	_____

(Adams and Frith 1979:38)

Total scores ranging from 16-99 were then converted into Proficiency Levels using Table IV below. This follows a scoring strategy in use since 1948 in the Defense Department with its Defense Language Proficiency Test (DLPT), earlier Army Language Proficiency Test (ALPT) (Petersen and Cartier 1975). It is curious that the factors led FSI, on the one hand, toward greater differentiation, adopting the Index score strategy after the factors, and DLI, on the other hand, toward less, abandoning the index in favor of the Speaking Performance Profile with its factors.

Table IV Conversion Table from Index Score to Proficiency Level (ca. 1967)

Score	Rating	(Steps in Range)
16-25	0+	10
26-32	1	7
33-42	1+	10
43-52	2	10
53-62	2+	10
63-72	3	10
73-82	3+	10
83-92	4	10
93-99	4+	7

(Adams and Frith 1979:38)

At first glance, this Index scale of 16-99 appears to be an 83-point scale. In fact, it is not. Because each factor level corresponds to a single weighted score in Table III, no factor has more than six points of differentiation. Furthermore, as seen in Table IV, there are no point values corresponding to the lowest and highest proficiency levels, 0 and 5. Wilds notes that those ratings are exceedingly rare, leaving a *de facto* 9-point proficiency scale. This means that the six points on the factor scale corresponded to nine points on the proficiency scale. As a result, they were commonly broadened by testers making marks between factor scale points (Wilds 1979:9).

By the early 1980s, the Index score, which was "on the books" as the scoring method, had fallen into widespread disuse. As part of his overall revision of testing, Argoff revived the Index score as a way to focus attention on the five factors (H. David Argoff, Personal Communication, February 22, 1994). When Thea C. Bruhn took over the Testing Unit in the mid-1980's, the manner of using Index and Factor to assign Proficiency Level became codified into the system in use today.

In its revised form, the new Index score goes to 120 and corresponds to the 0-5 ILR scale (Cornwell & Budzinski 1993ms:29). Raters first assign numeric scores on each of the factors. The five speaking factors currently in use are weighted differently from the original Index. In the new system, comprehension contributes a maximum of 25 Comprehension contributes a maximum of 25 points, Discourse 25, Structure 30, Lexicalization 30, and Fluency 10. Each of these factors can be rated at any score within that maximum, unlike the original weights which gave one score for each level. The result is greater differentiation overall, and especially on the more heavily weighted factors.

It is unclear whether the weights assigned in Table V (below) differ from those in Table III due to re-analysis or to philosophy. I find no record of any regression or discriminant analysis of the new factors in the assignment of their weights. Adams' (1980) discriminant analysis of the relative contribution of (old) factor scores to prediction of the global score suggested that the amount of contribution differs according to the proficiency level, and Clifford's (1980) work on the same project showed a distressingly low ability to predict the global score from the factor scores. That study should argue against a fixed weighting system. The new system is not a strict weighting system, however; instead it offers greater range in each of the factors, but each point has the same value in the final score.

3.2. Rating

At FSI, the move toward further differentiation was spurred in large part by efforts to improve rating validity, that is, to be sure that raters were paying full attention to all aspects of the descriptive criteria. An important cause of error in naive global ratings is the de facto consideration of only one or two factors (Adams 1980).

At FSI, the official position in rater training is currently that rating proceeds from the components to the global. This is a long way from the earlier position as noted by Wilds (1979:9): "Partly because the sample was based mainly on tests in Indo-European languages, partly because of a wide-spread initial suspicion of statistics among the staff, use of the scoring system has never been made compulsory or even urged, though the testers are required to complete the [Table I] Check List."

Staff suspicion of the statistical approach was actually well-founded. The imposition of a fixed-point additive system on a criterion-referenced evaluation of an interview test is fraught with complications. For example, if raters often marked between points on the early checklist, what point value should they have assigned in such a case?

The newer, current Index scale maps onto the proficiency scale somewhat differently, as in Table V below. This tentative proficiency rating must next be carefully matched to the published Proficiency Level descriptions. If the description fails to match the test sample, the appropriate level is selected and the partial index scores are reassessed until the total Index Score falls within the proper range. The evaluation process is thus recursive, beginning with the speaking factors, proceeding to the global proficiency, and cycling back until they agree.

Table V Conversion Table from Index Score to Proficiency Level (current)

Index Score	Proficiency Rating	(Steps in Range)
<5	0	5
6-18	0+	12
19-31	1	12
32-44	1+	12
45-57	2	12
58-70	2+	12
71-83	3	12
84-96	3+	12
97-109	4	12
110-118	4+	8
119-120	5	+2

(FSI Language Proficiency Report Worksheet/1354A)

3.3. Calibration

The current Index Score provides for greater differentiation within each of the levels, generally 13 points per level (Table V). That lends quantitative expression to such notions informally tossed about by federal language proficiency raters as "a strong 2." On the other hand, the implications of the Index Score are in conflict with assumptions about scale calibration. While agencies generally agree that the ILR scores map onto the underlying proficiency scale at expanding intervals, as noted above, the Index Score system imposes explicit equal intervals throughout the core of the scale. These intervals are more explicit even than the linear representations of the ILR Speaking Performance Profile, because of the additive nature of the partial index scores. Worse yet, the intervals of Index Scores equated to a proficiency level shrink at the top end of the scale, precisely where they are assumed to be the largest.

Another calibration problem concerns the weighting of the factors as they contribute to a simple additive Index. The current Index does not give true weights as the old one did. In the old Index, moving a notch in a particular factor meant adding or subtracting points depending on the weight. In the new scale, greater weight really means more total points per factor. Across factors, however, a point is still a point. That is, the 2 points of a low fluency rating each have the same weight in the sum total as each of the 22 points of a strong lexicalization rating.

In the original discriminant analysis, Adams showed, as pointed out again in Higgs & Clifford (1982:69) and Lowe (1988:74), that the factors do not contribute uniformly at each level of global proficiency. Thus, while vocabulary may have a major influence on proficiency in the lower levels where simpler sentence structure will be adequate, grammatical structure of a different degree of complexity is more likely to slow or halt progress at the higher levels where lexical work-arounds may go unnoticed. Lantolf and Frawley (1988:183) point out, for example, that advanced learners use more words to do things that would take both beginners and native speakers far fewer words. Because each point of the Index score contributes equally to the total, such possible differences are explicitly ignored.

4. THE PERFORMANCE LEVELS

While greater and greater differentiation is seductive because it suggests greater information, some of that information may be of questionable validity, particularly in federal testing which is fundamentally global, both in elicitation and rating. It can reach a point where what began as a heuristic becomes a hindrance.

4.1. Generalization

At one such point, FSI turned back in the direction of generality. In the early 1980's, H. David Argoff revised the factor definitions and developed six Performance Levels with descriptive criteria, the same number as the 0-5 Proficiency Levels, which had also been recently revised by Argoff, Lowe and Clark (1983). The rationale for the changes derived largely from the semantic theories of Leonard Talmy -- both in terms of the expansion of "vocabulary" to the broader "lexicalization" and also to the holistic treatment of factor rating that the Performance Levels re-introduced (H. David Argoff. Personal Communication, February 22, 1994).

4.2. Rating

The performance levels are intended as a first step in assigning a partial index score on a factor. They are: Blocking, Dysfunctional, Intrusive, Acceptable, Successful, Superior (Cornwell & Budzinski, 1993ms:49-50). The Performance Levels are criterion-referenced, as are the Proficiency Levels themselves; however, the descriptive criteria for the Performance Levels are more focused within the factors. For example, the description of comprehension in the 1 Proficiency Level description is just one sentence:

> "Misunderstandings are frequent, but the individual is able to ask for help and to verify comprehension of native speech in face-to-face interaction." (ILR, 1983) [4]

By contrast, the Dysfunctional Performance Level of the Interactive Comprehension factor reads as follows:

"The Examinee recognizes his/her own non-understandings, which are quite frequent, but may not exhibit comprehension even after repetition or clarification. The Examinee's Comprehension is generally limited to common topics on which the Examinee and the Tester share knowledge [The ILR Proficiency Level description of the S-1 level notes 'familiar topics.' - LHG] The native speaker [Tester] must adjust the speed and complexity of speech in order to be understood. In languages in which there is a considerable shared international vocabulary with English, the Examinee's comprehension may appear to be quite precocious, but this is not supported by evidence of comprehension of non-shared aspects of the language" (Cornwell and Budzinski 1993:170).

The level of detail is quite different, but also Performance Level and Proficiency Level do not correspond exactly. The S-1 Level description says nothing about whether the Examinee understands what is said after clarification. The Dysfunctional Performance Level, however, specifies that help does not always help. By contrast, at the Intrusive level, the examinee still asks for clarification, but the help is more effective.

Rather than a Bottom-Up, Prove-You-Can-Do-Something approach or a Top-Down, How-Low-Can-You-Go approach to determining levels, Argoff advocated a Start-from-the-Middle approach; that is, start with the assumption of Acceptable performance and show progressively how that test performance is less than Acceptable or exceeds that level.

Once the Performance Level has been chosen, the specific partial index score is chosen according to the scales given in Table VI.

Table VI Mapping FSI Index Scores onto Test Performance Levels

Performance Levels	Partial Index Scores					
Comprehension and Discourse Factors:						
Blocking	0	1	2	3	4	
Dysfunctional	5	6	7	8	9	
Intrusive	10	11	12	13	14	
Acceptable	15	16	17	18	19	
Successful	20	21	22	23	24	
Superior	25					
Structure and Lexicalization Factors:						
Blocking	0	1	2	3	4	5
Dysfunctional	6	7	8	9	10	11
Intrusive	12	13	14	15	16	17
Acceptable	18	19	20	21	22	23
Successful	24	25	26	27	28	29
Superior	30					
Fluency Factor:						
Blocking	0					
Dysfunctional	1	2	3			
Intrusive	4	5				
Acceptable	6	7	8			
Successful	9					
Superior	10					

(FSI Language Proficiency Report Worksheet / 1354A)

On the surface, it may appear that this conflation of partial index scores into six Performance Levels is a step toward the Speaking Performance Profile with its six proficiency levels, or the original FSI system of rating factors on the Proficiency Levels, but the similarity is more apparent than real. The FSI Performance Levels are part of an additive model; they are merely a heuristic for obtaining partial index scores which are then added to get a total Index Score, itself a heuristic.

This great attention to heuristics reflects great concern for supporting and guiding the raters; however, the question of complexity for raters also arises. In fact, although official FSI procedure is that component ratings lead to global ratings, it is clear that most experienced raters make tentative global ratings first. Indeed this step is essential for the testing team in gauging the elicitation strategies, including the selection of text category for the reading test (Frederick H. Jackson, personal communication, 2/15/94). This is consistent with findings that global ratings are cognitively more immediate, even if they subsequently lead to partial index scores which provide greater differentiation (Littlefield 1986, Braungart-Bloom 1986).

4.3. Calibration

A closer examination shows that the way in which the partial scores add up to the Index Score and convert to the Proficiency Level is partially compensatory. If the six Performance Levels were merely renamings of the 6 Proficiency Levels, then according to the 'minimum score model', minimum partial scores at the Intrusive level of each factor (10+10+12+12+4 in Table IV) should add up to the minimum Index score that maps onto the 2 Proficiency Level in Table V. Instead, the minimum "all-Intrusive" score of 48 is three points higher than the minimum Index Score for a 2. In other words, an examinee could score at the top of Dysfunctional in three of the five factors and still reach a 2 Proficiency Rating, at least mathematically. In reality, a final check of the proficiency level criteria would likely lead to an adjustment of the Index Score, which raises further questions on the calibration of the Performance Levels against the Index and the Proficiency Levels. In a more extreme case, a strong score in just one factor could compensate for weaker scores in all the other four factors.

Table VII, which maps the Proficiency Levels and the Performance Levels in terms of the Index Scores, show that the S-1 Proficiency Level is more generous, i.e. starts at a lower Index Score, than the Dysfunctional Performance Level. By contrast, both Proficiency Levels 2 and 3 extend three or four Index points beyond the Intrusive and Acceptable Performance Levels respectively; that is, one could have minimally Acceptable on three or four of the five factors and still not attain a 3 level in proficiency, the reverse of the situation at the Dysfunctional level described above.

Table VII Mapping FSI Performance and Proficiency Levels onto Index Scores

Performance Levels (Ranges if all factors are at same Performance Level)

Blocking	Dysfunctional	Intrusive	Acceptable	Successful	Superior
0-18	23-43	46-67	72-92	97-117	120
0-18	19-44	45-70	71-96	97-118	119-120
0/0+	1/1+	2/2+	3/3+	4/4+	5

Index Score ranges corresponding to Proficiency Levels

A further scale calibration problem arises in the equal-interval, additive way that the scores of each factor accumulate from one Performance Level to another. Young (1992) criticizes this monotonic approach to scaling all factors in the ACTLFL and UCLES (Univ. of Cambridge Local Examinations Syndicate) rating scales, precisely because they imply a linearity of presence or absence of features when that ordering may not necessarily exist, as in the acquisition of complex discourse functions. On the one hand, the FSI factors are careful to include discourse functions for evaluation; on the other, conclusions which devolve naturally from this level of specificity in the rating scales exceed their theoretical foundation.

It should be noted, too, in the move toward generalization, the basic levels and their plus-levels are conflated in Table VII, marking something of a return toward the notion that plus-levels are not entirely independent levels.

5. THE ROAD BACK TO GLOBAL RATING

To summarize so far, one overarching set of problems as rating has evolved has been concern for **the relative contribution of different aspects of proficiency to the overall rating**. The first approach to this problem was to set up a set of factors for consideration, the next was to weight these factors according to their contributions. Both the factors and the weighting system were later revised, bringing in the Index score concept with its greater range of differentiation among factor ratings. It might be possible, taking advantage of this differentiation, to utilize the partial index scores as an unofficial proficiency profile of the type recommended by Bachman.

On the other hand, despite its apparent support for the rater, the system is really more complex than that of the other agencies. Furthermore, the system of weighting and adding partial index scores in the various factors exceeds its theoretical basis. Indeed, it even exceeds its empirical validity; Adams' discriminant analysis showed that the factors contributed differently at the different proficiency levels, but the weightings did not reflect that. Clifford showed that the factor scores did a poor job of predicting global rating. Much has changed since Adams' study: the factors themselves, the index score system, and the proficiency levels. This summer, the Testing Unit and the Research, Evaluation and Development Unit at FSI will work together to replicate the analysis, to see whether these original observations held up in the current system.

The simpler, global-first approach to rating seems to be a cognitive inevitability. It is supported both by research with controlled comparisons and by the cumulative experience of decades of language testing and tester training in the federal government. Even the administrative use of a single global score per skill, interpreted in fact as a single total score, is evidence of a certain cognitive need for global-first.

A second overarching problem was that no system for representing these factors seems to take into account the expanding nature of the scale. This problem showed up in the measuring of centimeters along a line, score averaging, the adding of scores point by point, and the fixing of plus-levels.

FSI tester certification training makes no reference to a psychometric scale of any kind. Emphasis is overwhelmingly on the criterion-referenced quality of the rating; test samples are matched to level descriptions. Indeed the position of the FSI Testing Unit has been that the ILR scale can not claim to have any interval definition at all; it is purely ordinal (Thea Bruhn, personal communication, 12/92). Yet, as noted above in Step Two, the Index Score system adds up to explicit equal-interval units that are truncated at the scale extremes. The Index Score system is not the only one with this problem; the Speaking Performance Profile used by other Roundtable agencies, with its centimeter scale, is also equal-interval.

Neither the undefined nor the equal-interval models hold up, however, in the minds of the raters themselves. In a recent survey of FSI testers and examiners on how they perceive the intervals between the levels (Hart-Gonzalez 1993b), there was general agreement on expanding intervals as the appropriate model for the ILR Proficiency Scale in FSI testing. This is the model explicitly held by the rest of the Roundtable community.

A third problem was whether or not scores represent bands of proficiency evenly. In FSI, where scoring allows slight compensation among factors at all levels, proficiency scores represent essentially a midpoint in the proficiency ranges they represent. The Index Score strategy of FSI allows for compensation across all levels, all scores represent their respective bands of underlying proficiency in a uniform way, best thought of as the mid-range. This is in contrast to the CIA/DLI/FBI system which allows for some compensation among factors at the lower end, but not the upper end of the scale.

At this time, agencies of the Roundtable have been meeting on a Testing Board at the Center for the Advancement of Language Learning (CALL) to re-examine federal language testing, including among other things the rating criteria and procedures. These developments will surely be addressed in greater detail in the Federal Language Testing roundtable being presented at this 1994 Language Testing Research Colloquium (Cascallar et al.).

In brief, the agencies have agreed upon some things. For example, as noted before, the scale is best mapped onto the underlying proficiency range in expanding intervals. It is as yet unclear how plus-levels should be treated. As noted above, they are treated as separate levels in some agencies' administration. Their calibration in an expanding scale, however, is complicated, and sometimes misleading.

In terms of rating, several points of agreement were reached. To start, global rating happens first -- in fact, during the test process -- and it is a necessary part of the elicitation strategy. This is represents a slight shift in FSI position, but one already clearly in the minds of testers, who could not do otherwise. Second, rating aids (such as profiles, factor scores, etc.) are for use when raters are not confident that the global rating conforms fully to the

Proficiency Level criteria. In practice, this is most likely to happen around the plus-level/next-level boundaries.

Third, the rating aids (such as factor scores) are for the sole purpose of arriving at the global scores. And finally, the rating aids devised will be used by all agencies, and will be used the same way, as criterion-referenced, six-point factor scales. This removes the rating aids from the tangle of weightings and linearity. It also denies the possibility of using these factor scores as some sort of language learner profile.

In sum, the system which is now in development will provide greater guidance to the rater by elaborating criterion descriptions, not by imposing further steps in quantification. The removal of these extra levels also clears some of the unwanted assumptions of scale. The issue of plus-levels, however, remains.

NOTES

1. I wish to express my strongest appreciation for the cooperation, guidance, and assistance of my colleagues at this and other agencies, including Katrin Gardner and Pardee Lowe, Jr. (CIA) and John Clark, John Lett, Herb Davy, and Pat Dege (DLI) for their time, assistance and guidance. I also credit informative discussions with John Clark (DLI), Madeline E. Ehrman, Frederick H. Jackson, and David A. Argoff (FSI) and Claudia P. Wilds (formerly FSI) for uncovering interesting points of history. In addition, I wish to thank the members of the CALL Testing Board who reviewed some or all of this paper in draft form: Eduardo Cascallar, Julie Thornton, Katrin Gardner, Olga Navarrete, John L.D. Clark, Madeline Ehrman, Frederick Jackson, Sigrun Rockmaker, Stephen Sokolov, and Burt Weisman. With their help I have tried to capture agency testing fairly and accurately; to the extent that I fail, the shortcomings are all mine.

2. These components of the language proficiency ratings are commonly called "factors" in the federal language community. This is not in the statistical sense of the word, as these factors are not based on a factor analysis.

3. This Index Score is generally attributed to Claudia Wilds , but the 1979 Spanish and French Test Kit is mistakenly listed in her name by the Library of Congress. It was edited by Adams and Frith and came out long after Wilds had left FSI in 1972 and unbeknownst to her (Claudia P. Wilds, personal communication, 2/14/94). The material attributed to her in the volume, however, has an estimated date of 1967. This means that the weighting of the factors on an index score had already been instituted, and she mentions an R-score from a regression analysis in her 1975 presentation at Georgetown University (when the Interagency Language Roundtable preceded the Georgetown University Roundtable on Languages and Linguistics).

4. This passage also highlights the difference between interactive listening, evaluated here, and non-interactive listening.

REFERENCES

Adams, Marianne Lehr, "Five Cooccurring Factors in Speaking Proficiency." In Frith, James R. (Ed.), *Measuring Spoken Language Proficiency*. Washington, DC: Georgetown University Press, 1980:1-6.

Bachman, Lyle. "Test Usefulness: Principles and Consideration in Designing a Foreign Language Assessment System." Keynote paper and discussion at RP-ALLA 93: *Research Perspectives in Adult Language Learning and Acquisition – Articulation: The Role of Placement and Proficiency Testing*. Columbus, OH: Nov. 1993.

__________. Constructing measures and measuring constructs. In Harley, B. et al. (eds.) *The Development of Second Language Proficiency*. Cambridge: Cambridge Univ. Press, 1990, 26-38.

__________. Problems in Examining the Validity of the ACTFL Oral Proficiency Interview. *Studies in Second Language Acquisition* 10 (1988): 149-164.

Braungart-Bloom, Diane S. "Assessing Holistic Raters' Perceptions of Writing Qualities: An Examination of a Hierarchical Framework Following Pre-Post Training and Live Readings. Apr 1986.

Cascallar, Eduardo C., Cascallar, Marijke I., Child, James R., and Lowe, Pardee Jr. "Translation Proficiency Skill Level Descriptions: A Report of New Descriptors and the History of Their Development." Presented at the 27th annual Meeting of the American Council on the Teaching of Foreign Languages (ACTFL). San Antonio, TX: Nov. 1993.

Clark, John L.D. "A Study of the Comparability of Speaking Proficiency Interview Ratings Across Three Government Language Training Agencies." Washington, DC: Center for Applied Linguistics, 1986.

__________. *Direct Testing of Speaking Proficiency: Theory and Application*. Princeton, NJ: Educational Testing Service, 1978.

__________ & John Lett. "A Research Agenda." *Second Language Proficiency Assessment: Current Issues*. Ed. Pardee Lowe, Jr. & Charles W. Stansfield. Englewood Cliffs, NJ: Prentice Hall Regents, 1988:53-82.

Clifford, Ray T., "Foreign Service Institute Factor Scores and Global Ratings." Ed. James R. Frith, *Measuring Spoken Language Proficiency*. Washington, DC: Georgetown University Press, 1980:27-30.

__________, "Convergent and Discriminant Validation of Integrated and Unitary Language Skills: The Need for a Research Model." *The Construct Validation of Tests of Communicative Competence*. Ed. Adrian S. Palmer, Peter J.M. Groot, & George A. Trosper. Washington, DC:TESOL, 1981:62-70.

Cornwell, Isabella & Anna Budzinski. *FSI Language Tester Training Manual*. Ed. Thea Bruhn. Arlington, VA: Foreign Service Institute, March 1993 Pilot Copy.

Davies, Alan. *Principals of Language Testing*. Cambridge, MA: Basil Blackwell, 1990.

Frith, James R. (Ed.), *Measuring Spoken Language Proficiency*. Washington, DC: Georgetown Univerisyt Press, 1980.

Hart-Gonzalez, Lucinda. "Measurement Assessment Study: The Language Proficiency Scale." Paper, American Council on the Teaching of Foreign Languages, San Antonio, TX, November 1993.

__________. "Language Proficiency Testing in Federal Research." Syposium Paper. *Research Perspectives on Adult Language Learning and Acquisition*, Columbus, Ohio, October 1993.

Herzog, Martha. "Issues in Writing Proficiency Assessment, Section 1: The Government Scale." *Second Language Proficiency Assessment: Current Issues*, Ed. Pardee Lowe, Jr. & Charles W. Stansfield. Englewood Cliffs, NJ: Prentice Hall Regents, 1988:149-177.

Higgs, Theodore V. and Ray, Clifford. "The Push Toward Communication." *Curriculum, Competence, and the Foreign Langauge Teacher*, Ed. T.V. Higgs. Skokie, IL: National Textbook, 1982:57-79.

Jones, Randall L. and Spolsky, Bernard. *Testing Language Proficiency*. Washington, DC: Center for Applied Linguistics, 1975.

Klein-Braley, Christine. "Ask a Stupid Question...: Testing Language Proficiency in the Context of Research Studies." *Foreign Language Research in Cross-Cultural Perspective*. Ed. Kees de Bot, Ralph B. Ginsberg, and Claire Kramsch. Amsterdam: John Benjamins Publishing Co., 1991: 73-94.

Littlefield, John H. and Troendle, G. Roger. Lowe, Pardee, Jr. *Handbook of Question Types and their Use in LLC Oral Proficiency Tests* (Preliminary Version). LLC Internal Document, 1976.

Milanovic, Michael et al. (Eds.). "Developing Rating Scales for CASE: Theoretical Concerns and Analyses." Paper presented at the Annual Language Testing Research Colloquium, Vancouver, 1992, (ERIC Clearinghouse No. FL020187).

Petersen, Calvin R. and Cartier, Francis A. "Some Theoretical Problems and Practical Solutions in Proficiency Test Validity." *Testing Language Proficiency*. Ed. Jones, Randall L. and Bernard Spolsky. Washington, DC: Center for Applied Linguistics, 105-118.

Talmy, Leonard. "The Relation of Grammar to Cognition--A Synopsis." *Proceedings of TINLAP□ (Theoretical Issues in Natural Language Processing)*, 2. Ed. David Waltz. Urbana, IL: Univ. of Illinois.

Wilds, Claudia P. *Testing Kit, French and Spanish* ed. by Marianne Lehr Adams and James R. Frith. Washington, DC: Dept. of State Foreign Service Institute, 1979.

Young, Richard. Expert-Novice Differences in Oral Proficiency Testing. February, 1993.