DOCUMENT RESUME

ED 298 759 FL 017 546

AUTHOR Fulcher, Glenn
TITLE Lexis and Reality in Oral Evaluation.
PUB DATE 88
NOTE 62p.; Revised and expanded version of a paper presented at the Annual Meeting of the International Association of Teachers of English as a Foreign Language (22nd, Edinburgh, Scotland, April 11-14, 1988).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Information Analyses (070) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Discourse Analysis; Interviews; *Language Tests; *Rating Scales; Second Language Learning; Test Construction; *Test Validity; *Vocabulary Skills
IDENTIFIERS *Interagency Language Roundtable Oral Interview

ABSTRACT

A study investigated the rating scale used in the Interagency Language Roundtable (ILR) oral interview, focusing on the concept of vocabulary underlying assessment. The study examined the differences in strategies used by fluent native speakers and non-fluent non-native speakers to avoid disruption of communication in real conversation when lexical items are not available to them. It is argued that the ILR's concept of vocabulary is too vague to be of real practical value in an operational testing model, and that data-based discourse analysis techniques for test construction can be used to overcome the scale's shortcomings. (MSE)

Lexis and Reality in Oral Evaluation*

Glenn Fulcher

Director of English Studies
The English Institute
Nicosia, Cyprus.

1988

Abstract.

This paper looks at the development of an oral assessment scale in current use and the principles underlying it from the point of view of the assessment of the "vocabulary" of the testee. It argues that the concept of vocabulary underlying this scale is too vague to be of great practical value within an operational testing model, and suggests that the use of data-based discourse analysis techniques in test construction will lead to various developments in rating scales which could overcome the problems isolated.

---

* This paper is a revised and expanded version of a paper delivered at the 22nd International Conference of the International Association of Teachers of English as a Foreign Language and TESOL Scotland, at the University of Edinburgh, 11-14th April 1988. My thanks are due to Dr. J. Charles Alderson for scrutinising the first version of this paper and offering much good advice. I have not followed Dr. Alderson's suggestions at every point, and so any errors in argument, reference or expression remain my responsibility alone.

Getting Closer to Reality.

Any testing situation is likely to be "manipulative" in that the situation is created entirely for the purpose of testing. and both the tester and the testee will be aware of this (Spolsky, 1985, 34). Carroll suggests that one way of overcoming this problem is to use "non-invasive assessment": actually watching the testee carry out some real-life task in which the use of the L2 is essential (Carroll, 1985, 67-68). Here, the observer would be provided with assessment scales based on a needs analysis of the task stating the functional language requirements for any person to carry out the task effectively. This striving for "reality" has even led Spolsky (1985, 38) to consider the possibility of using the "planted encounter" where the testee would not be aware that he was being assessed at all.

The practical problems, the ethical problems of the latter approach. and the sheer cost of such a system all preclude its widespread use.

Testers are therefore thrown back upon the need to create situations from which they can attempt to predict success in

situations beyond that of the test. Prediction and not direct access (even if this is truly possible without sampling) seems set to remain the corner-stone of testing theory.

However, it is widely claimed that in oral assessment "communicative testing" has had its most obvious success because the situations used together with the assessment scale closely mirror real-life factors in spoken interaction (Carroll, 1985, 45; Lowe, 1987).

With reference to data-based criteria it has already been argued that the Carroll (1980) scales provide an inadequate definition of "fluency" when measured against samples of native speaker informal conversation (Fulcher, 1987a), and these conclusions are in agreement with Hieke (1985). This problem with the rating scales is that they are internally logical but have little reference to external reality (Lantolf and Frawley, 1985). However, what has not been considered is whether "real-life" is meant to be what native speakers can do (touched upon in Alderson, 1981), or whether it is to include what competent non-

native speakers can do. This second position seems to be implicit in Carroll's "functional profiles" (Carroll, 1978; 1980) but does not appear to have any direct effect upon the scales developed. Most of the assessment scales in current use, including the one discussed here, use the label "native" or "native-like" speech in the top band, thus implying that native competence is the yardstick upon which non-native performance is to be measured.

This presents a serious problem. Not the expected one of how a non-native speaker can approach native speaker "competence", but rather what criteria should be used at each level in the scale to state how a student falling into that band differs from the "ideal native speaker competence" and, more importantly, from a speaker who is placed in a band directly above or below. The key criterion of "hesitation" in the assessment of fluency seems to occur frequently in native speech (Fulcher, op. cit.) which raises the question of whether or not there are any observable linguistic signals associated with hesitation in non-native speech which would make it clear that the hesitation was in fact due to some element of language restriction. It appears to be the case that such linguistic signals are very often difficult to

isolate. For a rating scale which relies upon "hesitation" a a criterion of restricted language competence it does seem appropriate that the scale distinguish performance characteristics of the non-native speaker from performance characteristics in native speech where the native speaker is judged to be merely reformulating in real-time processing. [1]

The same point applies to the assessment of vocabulary, which is not unconnected to the assessment of fluency. The following fragment of conversation is taken from Crystal and Davy (1975, 19):

> "and he's been to America he's he's been to the la to oh the last f f two or three world cup world cup mat things you kno tournaments... "

If this had been a non-native speaker, the failure to find the word "tournaments" for "matches" and the mediating use of "things" (pervasive in "real" spoken English) would probably have resulted in the testee being marked down. In native speaker talk it is accepted unconditionally by the native listener who manages

to process the message effortlessly, and seems not to notice performance hesitations unless they are actually pointed out to him in transcript form after the event. Hesitation seems to be important in the real-time processing of language.

FSI to ILR: Speaking level Descriptions.

The Foreign Service Institute (FSI) oral rating scale is widely regarded as "the mother" of rating scales for assessing students in the oral interview. As an alternative to a global scale, it offered five analytic scales covering accent, grammar, vocabulary, fluency and comprehension. The vocabulary scale is given in full (Table 1).

This scale is not now used, but was the first in a long series of developments which has lead to the latest generation of oral interview scales of the ILR (Interagency Language Roundtable). The history of the development of these scales may be traced in Sollenberger (1978), Liskin-Gasparro (1981) and Lowe (1983).

Firstly, what comments may be made about the original FSI level descriptions for vocabulary? The first question is to what extent

Table 1: The FSI Vocabulary Sub-scale.

BAND	DESCRIPTION
1	Vocabulary inadequate for even the simplest conversation.
2	Vocabulary limited to basic personal and survival areas (time. transportation. food. family etc.)
3	Choice of words sometimes inaccurate. limitations of vocabulary prevent discussion of some common professional and social topics.
4	Professional vocabulary adequate to discuss special interests: general vocabulary permits discussion of non-technical subject with some circumlocutions.
5	Professional vocabulary broad and precise: general vocabulary adequate to cope with complex practical problems and varied social situations.
6	Vocabulary apparently as accurate and extensive as that of an educated native speaker.

the vocabulary scale can really be kept separate from the notion of fluency in general within the early FSI model For example. band 4 on the fluency scale reads: "Speech is occasionally hesitant. with some unevenness caused by rephrasing and groping for words". Apart from the problem of defining what phenomena may

constitute "unevenness", on what linguistic criteria should one distinguish "rephrasing" and "groping for words"? Is this, in practice, any different from "circumlocution"? Secondly, the example of native speaker talk from Crystal and Davy could very well be described using such terms as "hesitation", "groping" and "circumlocution" which would presumably place this educated native speaker within band 4 on both scales. This, of course, should not be the case, but occurs because native speaker talk has been idealised by the scale. This could explain why Jones discovered that the FSI often failed to discriminate after the level 3+ (Jones, 1985, 82).

The ILR scale does not offer separate analytic scales, but only one global scale, and thus avoids the problem of trying to distinguish skills in a conceptually and empirically valid way despite the lack of evidence to support the theoretical validity of such global scales since the demise of the strong version of the Unitary Competence Hypothesis. As such, in order to compare it with the original FSI vocabulary scale, references to vocabulary have been extracted and information adapted to table

form for the following analysis based on information in the Foreign Service Institute (ILR) descriptions (1985).

Table 2: The ILR approach to vocabulary.

BAND	DESCRIPTION
0	Oral production is limited to occasional isolated words.
0+	The individual's vocabulary is usually limited to areas of immediate survival needs.
1	Vocabulary is inaccurate, and its range is very narrow...speakers at this level may have encountered quite different vocabulary areas....Vocabulary is extremely limited and characteristically does not include modifiers....Use of...vocabulary is highly imprecise.
1+	Speech largely consists of a series of short, discrete utterances.
2	Vocabulary is appropriate for high-frequency utterances, but unusual or imprecise elsewhere.
2+	He/She is generally strong in either structural precision or vocabulary, but not in both....Normally controls, but cannot always easily produce vocabulary.
3	Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations on practical, social and professional topics....The individual can effectively combine structure and vocabulary to convey his/her meaning accurately....Without searching for words and phrases, the individual uses the language

clearly and relatively naturally to elaborate concepts freely and make ideas easily understandable to native speakers.

3+ Typically there is particular strength in fluency and one or more, but not all, of the following: breadth of lexicon, including low- and mid-frequency items....

4 (No specific reference to vocabulary)

4+ The individual has a sophisticated control of vocabulary and phrasing that is rarely imprecise, yet there are occasional weaknesses in idioms, colloquialisms, pronunciation, cultural reference....

5 ....speech is on all levels is fully accepted by well-educated native speakers in all its features, including breadth of vocabulary and idiom, colloquialisms, and pertinent cultural references.

A comparison of the two scales demonstrates that the concepts involved are very similar in many respects. We concur with Hieke's statement (1985, 137) that such tests are not "fair"

> "as long as they hinge upon such pros statements to delineate levels while these are peppered with notions that cannot withstand close scrutiny."

When terms and descriptions are unspecific and ill defined, interpretation of the scale is bound to be "relative" (Skehan, 1984, 217). [2] In particular, the ILR description has introduced

the term "precise" and "precision" without any indication of what this implies. A further oddity, certainly from the point of view of discourse, is how the constructors of the scale know that at level 3+ there is strength in "fluency" and one but not all of a list from "vocabulary", "structural precision" and "discourse competence" (= "a native-speaker's strategic and organisational abilities and expectations"). Only on one point is Lowe more helpful when he says that native-speakers typically fall at band 3 (Lowe, 1987), but this is a claim (presumably like the claims in the actual descriptors) which is made on the basis of "experience" and not hard empirical evidence. This leaves totally open the question of what a "well-educated native speaker" (band 5) might be: a point raised by Bachman and Savignon (1986, 383) in relation to the development of the ILR for use in the American Council on the Teaching of Foreign Languages (ACTFL) speaking level descriptions [3]

Similar criterial notions underpin the FSI and the ILR scales, including hesitation, groping for words, circumlocution, limited vocabulary, and educated native speaker proficiency (which does

not involve any of these things) being the ultimate goal of the learner. Some attempt is made to link these notions to what the testee can do. but in an intuitive way. through the descriptions of elicitation procedures. which may well be misleading when actual data is brought into the picture [4]

Both the FSI and the ILR scales range from n zero proficiency to native proficiency. and the levels between these two extremes view vocabulary primarily as "knowledge" (see for example, Adams, 1980. 4). One has "enough vocabulary" to carry out certain tasks, or a "limited vocabulary". Hesitation occurs when there is a gap in lexical knowledge. and circumlocution or groping signal such a gap. This lays bare the implicit theoretical underpinning of the ILR scale. which has not changed from the time of the original FSI. that from observing performance in testing situations one can make judgements about underlying competence, and the notion of competence - whether it be linguistic or communicative - relates to a theory of how learners progress from a zero state to a well-educated native speaker state, where lexical gaps, hesitations, repetitions and the like are finally eradicated from the observable performance.

Related to such theory Carroll discusses a major problem faced by the tester, that "he cannot test competence in any direct sense; he can measure only through the manifestations of it in performance" (Carroll, 1968, 51). All testing involves the use of theory, and all theory involves abstraction, but the abstraction must be empirically demonstrated to relate to the data on which it was developed. If this cannot be done then the test cannot be said to be valid. The data from discourse analytic studies has shown that the notion of an ideal native-speaker competence or performance is not realistic, and is breaking down in favour of the view that all that actually exists is a set of varying performances influenced by many contextual, textual and personal factors. For the competence/performance division to be maintained in any form which is useful to testers these factors must be taken into account.

As such, the view of vocabulary as "knowledge" in the competence of the learner needs to be reviewed. Many testers have said that the selection of lexical items for testing in oral examinations may be done on the grounds of frequency (see, for example,

Spolsky, 1986, 150 and Jones, 1981, 100) which must be elicited by the interviewer. This seems to be a far too simplistic approach which does not take into account the kinds of factors mentioned.

If, on the other hand, the tester concentrates first of all on the complexity of variety within performance he will become more concerned with strategies open to non-native speakers when a word is not available to them in the process of real-time conversation. The present scales do not take this into account. Two interrelated points arise from this discussion: (1) if a non-native speaker does not overtly signal that a word is not available to him/her then he/she may be behaving linguistically like a native speaker when hesitating etc., and cannot be penalised for this in the scoring, and (2) when a non native speaker does overtly signal that a word is not available to him/her it may be that the strategies employed by the "fluent" speaker do not disrupt communication and are seen as acceptable strategies by the listeners. In this case, the non-native speaker should not be penalised in the scoring.

When looking at non-native performance the question to be asked is: what are the differences in strategies used by fluent non-native speakers (FNNS) and non-fluent non-native speakers (NNS)? Comparison of performance strategies may well lead to the possibility of writing descriptions for levels within banding systems that actually reflect observed differences rather than hypothesised differences.

The Data.

In the data used for this study all non-native speakers were in an informal environment, but where there was some pressure on them to perform well because of the presence of native speakers who were meeting them for the first time, and a teacher of English who spoke their L1. Various other speakers of the L1 were also present on both occasions. The taped conversations last for a total of approximately 5.5 hours. These tapes were not all transcribed; they were played to a number of English teachers with some experience of oral examining who were asked to make notes on those parts of the tapes where it was thought that there were vocabulary problems in the recorded speech. This was done

from the first listening only.

As performance was the target of the study, an arbitrary definition of a fluent non-native speaker was a person who had completed tertiary education in an English medium university, while a non-fluent non-native speaker was defined as a person who was studying English in the two years prior to admission to an English medium university. The L1 of all speakers was Greek.

Fluent Non-Native Speaker Strategies.

For the most part, those falling into this category cannot be distinguished easily from native speakers from transcripts, and when listening to tapes accent is the factor which makes the difference. However, when it appears clear that a performance error which a native speaker would not make has occurred, it is signalled by the performance strategies used to avoid potential problems in communication. These strategies may be summarised as follows:

Table 3: FNNS Strategies.

PRODUCTIVE MODE (PM)	RECEPTIVE MODE (RM)
(a) Equivalent substitution	(a) Incomplete repetition
(b) Direct questioning (+ repetition)	

Firstly, it needs to be stated that this is not seen to be an exhaustive classification of FNNS strategies, merely those observed in some three hours of recorded data. The same is true for the NNS strategies. [5] Secondly, there appears to be a relationship between PM (a) and PM (b) which could be stated in the form of a tactical performance rule: if an equivalent word is available then use it: if not, then ask directly. (For the notion of equivalence, see McCarthy, 1988b.) The data will not support complete justification of this, but the plausibility that such tactical performance rules exist should be the object of further research with larger bodies of data.

In the RM, that is listening to another speaker and failing to understand a word from context which is important to the next utterance after a turn change, the only strategy evidenced in the data was incomplete repetition. For example, (NS = native speaker):

RM (a)

(NS) Is there erm a limit to the amount of goods you can take back into [name of country]

(FNNS) The amount of

(NS) goods you know apart from erm alcohol and perfume you know if you're taking back clothes and presents and things like that do you

(FNNS) well I suppose

(NS) know what the limit is

The FNNS repeats the nominal group with the exception of the the item which is causing the performance problem, and this is then explained by the NS by providing, in the existential category of inclusion, "clothes" and "presents", which resolves the performance problem within context. It should be noted that the items given in explanation are existentially included in the category of "goods" but not necessarily semantically, as meaning is being negotiated in context rather than in an abstract description of the language (see Brazil, 1985, 41; Carter & McCarthy, 1988, 212; Cruse, 1975, 29-30. These notions are similar to Hasan's (1984; 1985) concept of "instantial relations" in lexis in text.) The use and comprehension of such explanations demonstrates the existence of underlying linguistic and sociolinguistic competences. It may also be noted that this strategy is commonly used by native speakers when they either do not hear or cannot understand the "meaning" of a word in the PM. However, the NS may also use a Wh- word instead of a space: "The amount of what?" (McCarthy, personal communication.) This variation of the strategy did not occur in this data. FNNS strategy contrasts with NNS strategy on this same point in that they will either produce the Wh- word in isolation (see FNNS

data, RM (a)), or produce the characteristic "How do you call...", which is how the data in NNS PM (d) (ii) would probably have come out had the question been in English and not Greek. In the PM the FNNS will use equivalent substitution more frequently than direct questioning, as in this example:

PM (a)

(FNNS) well personally I mean he doesn't want to do the same thing although I think he should //p esPECially with that BLAzer he has//

(NS) // YEAH //p that CORduroy COAT // you mean you mean yeah well....

(FNNS) hmm

(Intonational information is included only where it is thought to be relevant to the interpretation of the example. In these cases the notation follows Brazil, 1985.)

"Blazer" is selected (and therefore given prominence in intonation) by the FNNS because it is communicatively appropriate and is seen as existentially equivalent to "corduroy coat", and the NS is able to redefine the terms being used (redefinition taking place through the use of high key) because the garment under discussion has been seen by both speakers. This particular strategy may therefore not be picked up at all by an interviewer if s/he is not familiar with the background information of the topic being discussed.

It is interesting to note in this particular conversation that the NS also redefines her own initial redefinition, because later in the conversation it is altered in the following way:

(NS) //p that corduroy JACKet //p was looking SHAbby....

This continual definition and redefinition of terms goes on continuously in all conversation, whether the speakers are native or not. In political debate where success depends on establishing

word meaning within an ideological context the phenomenon is much more prominent and may easily be observed.

Finally, if no equivalent item is available, then a direct question will be used. This may occur with or without repetition of the answer.

PM (b)

(FNNS) ....so they got rid of the green things that were coming on top and changed the skirt with some err what do you call the //r PLEATS // in the middle

(NS) //p PLEATS //r PLEATS

(FNNS) it was quite nice so I I wore it again....

(NS) hmm yeah

(A "p" tone is falling an indicates that new information is being introduced in this lexical item. As such, it is proclaiming rather than referring, while an "r" tone is a fall-rise, and indicates that the lexical item is referring to something that is

already "in-play" in the conversation.)

All these strategies appear to be reasonable non-disruptive performance strategies to maintain the flow of conversation, and differ from the vague "hesitation" criteria mentioned in the assessment scales. They do not give the impression of being "unnatural", and so should be classed as acceptable and therefore not penalised in scoring.

This makes it clear that data from non-native speakers is needed in the construction of oral assessment scales so that discrimination between performance levels, when made, is as valid as possible.

Non-Fluent Non-Native Speaker Strategies.

In this data the NNS strategies were extremely marked as different from those of the FNNS speakers. These strategies may be summarised in Table 4.

Table 4: NNS Strategies.

PRODUCTIVE MODE (PM)	RECEPTIVE MODE (RM)
(a) Communicative breakdown	(a) Direct questioning
(b) Substitution of L1 item	
(c) Checking by translation	
(d) Asking for translation of an L1 item.	

Whereas in FNNS data there appeared to be a hierarchical preference in the two PM strategies, no such relationship was observed in the NNS data, and the ordering given here is arbitrary. The categories require little initial discussion, but examples from the data will be given. All L1 items are transliterated in [square brackets].

PM (a)

(NNS) ....it's a stick erm this one it it's a stick over here (pointing at picture) and the rock was erm oh (laughs) I want you to say this please (laughs) oh come on

(NS) ah ah he rubbed the stick....

PM (b) (i)

(NNS) ....their animals were afraid they were running so here is erm what erm tch erm [kremos] you know

(ALL) (laughter)

(NNS) (laughs) and well I don't know how to say it there's the only way to understand it (laughs) and err so they....

PM (b) (ii)

(NNS) ....the first human beings were having their erm food from nature I mean that they didn't erm they didn't have any erm [sitira]. [Sitira] [sitira] Okay. They didn't work the land to give them err they were hunting....

PM (c)

(NNS) ....//p and adamski underSTOOD that it was erm erm //p VENus VENus //p //r [aphroDITe] //p VENus //p YES //r and he understood that they came from VENus //....

PM (d) (i): Question in L1.

(NNS¹) ....these people are far from our technology they are very erm what do we say erm [exelegmeni]

(NNS²) advanced

(NNS¹) yes

yes so just imagine that they used radiation and magnetic power....

PM (d) (ii): Question in L2.

(NNS¹) ....were there any witnesses

(NNS²) witnesses yes erm [pos to lene to vosko] (laughs) a man

(NS) a shepherd

(NNS²) yes yes but....

RM (a)

(NNS) ....are you thinking about something

(NS) yeah erm I I think I'm a sceptic about scep erm an unbeliever

(NNS) what oh

It is the use of L1 lexical items which is the most striking difference between the data from FNNS and NNS in comparable circumstances when other L1 speakers (natives and non-natives)

are present. [6] In NNS data there is an apparent lack of appropriate strategies to allow the conversation to continue without focusing directly on the language rather than the message; as such there is a change of plane (see Sinclair, 1983, for this concept) towards metalanguage and back again into the discourse may turn out to be a significant signal of reduced performance fluency due to lexical factors. The NNS, according to this data, appear not to see lexis in existential terms, and so do not use strategies open to the FNNS. Rather, NNS speakers seem to view lexis as something which is known or not known. It is claimed that this may be one explanation for the dependence on translation and the underlying assumption of a one to one semantic equivalence between L1 and L2 lexical items in NNS speech. ("Knowing" a word in traditional approaches to lexis involves knowing frequency of occurrence, collocability, limitations of use according to function and situation, syntactic behaviour, derivations, composition and polysemy (Richards, 1976; for more recent discussions see Beheydt (1987) and Anderson & Freebody (1981)). These are all clearly factors in an abstract system or systematic description of language, rather than the factors in real-time discourse processing discussed here. It is

not the purpose of this paper to discuss what it means to say that a learner know the meaning of a word - an enormous task - but to point out that there is a great deal of indeterminacy in this area which often does not relate to observable performance, which is the only way of assessing vocabulary in an oral testing situation.)

In the FNNS data the discourse is not disrupted by a plane change. That is there is no need to concentrate on the meaning of the lexical item before the conversation can continue, but the meaning is established existentially and contextually in the on-going conversation. In NNS data, however, there is a plane change to allow a quite deliberate focusing on the language rather than the message

If no L1 speakers had been present during these conversations it is to be suspected that the switch to the metalanguage would either have been a more drawn-out affair, or that a topic change would have followed a complete breakdown in communication. Whilst this would clearly be important to an interview situation where

the interviewer did not speak the L1 or the testee did not know that the interviewer spoke the L1, no data is available at present to substantiate such speculation. Nevertheless, the results of this study are still relevant in that they do demonstrate the existence of different performance strategies between FNNS and NNS which can be clarified for other situations in further research.

Comparisons and discussion.

The only "overlap" category between FNNS and NNS is that of direct questioning. However, in the recorded data all direct questioning in the FNNS speech occurs in the productive mode while all examples from NNS speech occur in the receptive mode.

It may at this stage be tentatively speculated that the NNS is more heavily dependent upon "classroom-type" strategies and has not yet made the leap to "real-life" strategies. The distinction here is, at this moment in time, purely intuitive, although it is not an unreasonable assumption to make Sinclair and Coulthard (1975) and Sinclair and Brazil (1982) have demonstrated the probability that patterns of interaction within the classroom are

different from other kinds of discourse. It may also be the case that lexical strategies differ too. If this could be firmly established through the analysis of data, then it would be possible to test whether or not classroom-type strategies rely heavily on plane changes within the discourse so that lexical items are discussed and seen by the students as having a one-to-one correspondence with an L1 item within a similar abstract semantic system. The most popular method of teaching lexis, giving a list of words with a gloss as they are encountered sequentially in texts may very well encourage this (James, 1985; Palmberg, 1986, Fulcher, 1987b). Most practising teachers of EFL have encountered the "rapid search" technique of students to discover the appropriate L1 item after the teacher gives the gloss in the L2. On the other hand, the "real-life" strategies used by the FNNS betray a state of mind which accepts contextual definition in the real-time discourse process, and the setting up of transitory existential categories. This factor may account for the feeling that the FNNS speech is more "natural" while the NNS speech is seen as in some way "artificial".

The Implications for Testing.

(a) For testing it should in principle be possible to develop a rating scale where the bands represent varying levels or types or performance. This would require a larger database than has been used here, as it can only really be claimed that these samples represent two kinds of performance. There may be other types between these two, and there will certainly be types of performance less fluent than the arbitrary category of NNS.

(b) The data-based discourse approach to constructing new rating scales hopefully avoids the problem of seeing many conversational phenomena as error in non-native speech while in native speech the same phenomena are not noticed. The higher bands of the scale are thus opened up to the non-native speakers, as they would no longer reflect an (intuitive) theory of native-speaker competence, but rather a high level of communicative performance. It is this problem that may had led to Jones' (1985, 82) comments on the FSI, and why Carroll (1967) found the cut off point on the older scales to be approximately 2/2+. Building on the ILR, the authors of the ACTFL scales have abandoned all bands above 2+ and just labelled them "superior" (Liskin-Gasparro, 1984). Rather

than opening up the higher bands, it has been argued that as they are so difficult to achieve, it is better to have many more bands at the lower end of the scale; whether this makes the test "more sensitive" at those levels, is another issue entirely.

(c) In the kind of scale suggested here there would be a fundamental shift away from discussions about how to sample the lexicon (Lado, 1978) towards an assessment method which can operate in real time without the need to consider the problems of word counts. The arguments put forward for the use of certain techniques such as the use of pictures to elicit certain lexical items begin to lose much of their force.

(c) If the scales are based on a large appropriate database the problem of the unequal status of the tester and the testee may be overcome to some degree. In the data for this study all non-native speakers were in unequal encounter situations despite the attempt to create an informal, relaxed atmosphere for the talk to take place. It does not seem unreasonable to argue that such unequal encounters are in any sense more "unreal" than the ideals

set out by Morrow (1979). They occur in every walk of life, not least between the tutor/supervisor and the student at university.

(e) Finally, with this approach the much discussed unreliability of assessors judgements may also be overcome to some degree by more careful definitions of the bands in the scale. Once it can be stated how one band differs from another above and below it in precise linguistic and sociolinguistic terms then problems in the scoring of oral tests (Jones, 1981) would be lessened.

Validation Procedures.

In any testing situation it is essential that the test is valid (for an explanation of the expression, as used here, see Palmer and Groot, 1981). Face validity is clearly important if the test is to be generally acceptable to both testees and testers, but it is important to recognise that this alone is not a sufficient criterion of a good test, and no amount of vitriolic cynicism directed at testing research (as in Underhill, 1987, 4-5) will make this the case. Nor, from such a standpoint, can the merging of content and construct validity be allowed to pass with criticism (ibid., 106). Validity in oral testing is a serious

issue which must be dealt with on the basis of arguments and evidence, rather than personal attack.

At present it is common for oral tests to be validated by reference to some criterion measure. Discussing a common measure of speaking proficiency, Clark (1980, 19) claims that

> "This requirement poses a major theoretical and practical problem in the development of the...instrument because there do not exist, at the present time, any sufficiently accurate or extensive criterion measures of real life communicative performance against which the...test could be validated."

He suggests the use of testee self-reports and independent non-intrusive assessment techniques to be used with the sample for validation to overcome this problem. More research is needed on the first method, but may be used as one of a number of validation techniques. The second method would be exceptionally time consuming and expensive.

Criterion (or concurrent) validity still remains the most popular

method of validating new scales. Ingram (1984, 18-19) points most lucidly to the fact that as most tests are based on different theories, using correlations with already available tests is a highly suspect business. This is, of course, not to say that there are no coherent defences of concurrent validation, for there certainly are (Davies, 1983).

However, it is suggested here that a data-based approach to an oral vocabulary scale based on strategies can open up a new approach to validation because the level descriptions would not be based on the intuition of the tester but on the observation of communication. Of course, no data remains simply as data. This would be to say nothing about it. As soon as it is analysed abstraction takes place, but fortunately applied linguists do agree that some abstractions are more likely than others, and abstractions based on data are much easier to validate by comparing them with bodies of data other than that on which the original abstraction was made.

In other words, construct validity rather than concurrent validity is to be given priority. In this view, the validity of

the oral rating scale is a function of the relationship in a (minimally) three-stage process:

Table 5: Scale development.

Observation¹	---------> Abstraction	---------> Observation²
	Relation¹	Relation²

The second relationship is that of validity. This relationship being satisfactory could lead to scale use.

Table 6: Scale use.

Observationⁿ	----------> Banding
	Relation³

Relationship 3 simply states that the band into which the student is placed is directly related to the observation of his performance compared with all original observations as described in the abstraction: the scale is a valid part of the testing procedure when the phenomena tapped by the scale correspond to the phenomena which the band descriptor describes. (It may be added that this approach acknowledges the fact that it is impossible to have criteria without norms!)

Although the stress on construct validity throws the emphasis in test development onto the analysis and description of data, the tester is not in any way absolved of the responsibility of providing empirical evidence to validate the construct statistically where possible and appropriate (See Standards 1.8, 1.9 and 1.10, in the APA Standards for educational and psychological testing). This is extremely difficult with oral tests, but one approach must be mentioned as it has great potential in this area.

The use of the multitrait-multimethod matrix (MTMM) has much to

recommend it (for a general introduction, see Henning, 1987, 101-105), and has been applied to the FSI oral rating scale by Bachman and Palmer (1983) among other tests. The MTMM approach as originally conceived by Campbell and Fiske (1959) fulfils two criteria which should be met in all validation studies if the results are to be interpretable: (a) the researcher must state in advance the type and nature of evidence which would cause him to reject his hypothesis or claim, and (b) the research technique used must provide a reasonable probability that the hypothesis will be rejected if it is untrue and accepted if it is true.

In particular, the MTMM approach allows the researcher to quantify the degree of influence upon results which is caused by test method factors, and in oral tests the elicitation techniques seem likely to be important variables which influence scores (Bachman and Savignon, 1986). Secondly, the approach provides a way of establishing whether or not various "sub-skills" (more often referred to as "traits", such as vocabulary, grammar, or whatever the theory upon which the test is based sets out in advance) are really conceptually and empirically distinct. Similar traits are tested for "convergence" and different traits

tested for "divergence" (Stevenson, 1981). The two aspects of test method factors and trait analysis can lead to the more accurate development of oral tests based on theory which ems directly from discourse analysis. In tandem, these two approaches may very well succeed in bringing construct validity to the fore of testing debates.

Conclusions.

It has been argued that the notion of reality as a criterion for evaluation within oral testing theory can be achieved, but that in the current scales "reality" is interpreted in relation to an ideal native speaker's competence rather than in relation to the observation of variety within the performance of non-native speakers. As such, the view of vocabulary presented by the scales described is "lexis as stored and retrievable semantic knowledge." The testing problem is analogously one of sampling and elicitation.

The alternative method offered here is a performance scale constructed upon data which reveal performance strategies. The

more successful strategies seem to be highly associated in the FNNS with a rejection of the "lexis as knowledge" position, and an acceptance of negotiated meaning in context.

A new oral rating scale based on discourse analysis (and the underlying constructs by implication) would need to be validated, and the criteria would be the degree of match between the construct and new data, the place of test method factors, and the degree of independence of hypothesised traits, as well as the relationship of the scale to external measures which rely upon a similar approach to testing. If, as Berkoff (1985, 96) has suggested, the key to problems in oral testing is how the bands of the rating scale are defined and how the criteria for rating are established, then this approach should in principle increase the validity of oral tests. If all raters are then properly trained in the use of the scale two of the most damaging factors associated with present scales may be overcome: the existence of a halo effect (factors external to the scale influencing the rating) and the central tendency error (placing students in the "average" band, and moving him/her up or down one or two bands depending on impression). Both of these factors are associated

with vague discriptors and/or poorly trained raters (Cronbach, 1984, 509-513).

Since 1978 when Davies noted that little work had been done on the application of discourse analysis to testing there has been a steadily growing stream of studies in this area. Future research must concentrate initially on linguistic and sociolinguistic description.[7] Having decided on the target population for a possible test, future research should then decide on how many levels/bands it wishes to have in its scale. Each band must then be associated with a subgro p of the population and the database established which can claim to be a fair representation of the discourse abilities of the subgroup. (In reality, the data will probably dictate the number of levels once the study begins, as data cannot usually be forced into a priori categories.) In describing each level, the tester must be sure that he concentrates on those aspects of discourse which do actually separate the groups from each other; the moment the prose descriptions become "squishy" (Hieke's term), then he should begin to suspect that the description is not adequate.

Once the descriptions are established, they must be validated against new banks of data. and elicitation techniques developed for the instrument: this is a complicated and important process in itself. but is not the purpose of the present study. Once the test is presented in a form that can actually be used with testee, traits must be validated and test method factors accounted for. Only at this stage can the oral test become operational on a wide scale.

The first stage in the process is that of description. This has been begun by some researchers (Perrett, 1987), but is really only in the early stages. In itself, this process shows that testing is. and should be. a two edged weapon. On the one hand it is used for assessment. but on the other its demand for operational models and validation provide both a tool and a check for researchers. In the next decade a discourse approach to testing may very well have much more of value to offer on both counts.

Footnotes.

1. The appeal to notions such as "hesitation" has a seductive appeal on the surface, but the use of such terms alone is inadequate because they are based on popular views of validity and not the more stringent demands of construct validity. Butterworth (1975, 76) reports on data-based studies in which it was

> "...hypothesized that the amount of speech in the fluent phase required the planning time given by the pausing in the hesitant phase, and [the studies] found that hesitant phases exhibited not only a greater proportion of pausing but also more hesitation phenomena of other kinds; for example there were more filled pauses - "ah's" and "um's"."

The psychological correlate is that there may be a relationship between quantity of speech/utterance length and "cognitive processing time" needed. Directly related to the present concern with vocabulary, in his study, Butterworth (ibid., 84) claims that

> "...difficult lexical choices have been revealed as meriting sufficient of the speaker's planning time to cause a delay in his output."

As such, it would seem that the phenomenon of hesitation has a psycholinguistic role in the vocabulary selection process of native speakers, and we are therefore not at liberty to assume that in non-native speech this phenomenon is a signal of language restriction of "limited competence".

2. It could be claimed by test developers that the level of descriptions are imprecise because of necessity, that they are a shorthand for the examiner which do in fact reflect clearly specified constructs. Hieke (1985, 137) concedes that for an "experienced evaluator" the terminology may be meaningful, but correctly goes on to say

> "...but to explain what...these terms mean at, say point four on the scale would stymie even the more grizzled among the raters."

The point is that raters tend to interpret the scales in the light of their own experience. However, the scales must represent constructs which can not only be validated, but have a database upon which differences between levels can be demonstrated to exist within the population to be tested. This is the old problem of not being able to describe performance adequately, and for progress to be made "definitions of performance must go far beyond the "you know what I mean" level." (Stevenson, 1981, 43).

3. The use of a scale which relies on the "native speaker" criterion as the yardstick for measurement has been severely criticised by Bachman and Savignon (1986, 383) because of the "considerable variation in ability" demonstrated by native speakers. (This also, they f l, the reason why recent scales based on the FSI rely upon similar underlying notions.) Spolsky (1973, 172) distinguished competence and performance in the following way:

> "One is said to know a language when one's native competence is like that of a native speaker. Performance need not however be identical, for it is accepted that someone knows a language even when he speaks hesitantly, with many errors, or with a foreign accent, or when he understands it with some difficulty under conditions of noise."

This is true of native and non-natives alike: it is only in foreign language testing that some have taken non-native performance errors to directly reflect competence. In interlanguage studies it has been recogni ed that "free elicited L1 performance...is anything but "fluent" in the sense of "ideal delivery of speech." (Raupach, 1983, 206). Differences in performance strategies between L1 and (levels within) L2 performance would help to begin to clarify the issues involved in the creation of yardsticks for testing, and throw light on the increasingly illusive notion of "competence".

4. The FSI and the ILR (like many oral tests) are said to be valid because of (a) the "directness" of the test and (b) because people have had a lot of experience in using it (Stevenson, 1981, 49; Bachman and Savignon, 1986, 382; Lowe, 1987). This claim rests on the face validity of the test, which is not an

acceptable criterion on its own for claiming validity

5. No taxonomy of strategies could possibly be anything more than transitory given what is currently known about such aspects of communication. That provided by Blum-Kulka and Levenston (1978) was based on data from Hebrew, but their strict adherence to semantic categories in their group A, and in group B the lack of importance attached to contextual/situational strategies (which they acknowledge on page 137) means that it is now not adequate for data description.

6. This is what I take to be virtually equivalent to the notion of "Language Switch" in data-based interlanguage studies (Bialystok, 1983, 105).

7. The ability to define the performance domain which is essential to the validation of oral tests must deal with linguistic elements, sociolinguistic elements, and their interaction in the testing situation. In the construction of the test and its validation, elicitation procedures must not be assumed to be the same as the constructs to be evaluated, for to

do so would be to confound method and trait - exactly what the MTMM is designed to avoid. This is one reason why a focus on strategies as indicative of psycholinguistic processes and sociolinguistic competences may have a very practical benefit when it comes to designing validation studies.

Bibliography.

Adams, M. L. (1980) "Five cooccurring factors in speaking proficiency" in J. R. Frith (ed) Measuring Spoken Language Proficiency. Georgetown University Press, 1-6.

Alderson, C. J. and Hughes. A. (1981) Issues in Language Testing. ELT documents 111, The British Council.

American Psychological Association (1985) Standards for educational and psychological testing. APA.

Anderson. R. C. and Freebody. P. (1981) "Vocabulary Knowledge" in T. Guthrie (ed) Comprehension and Teaching: Research Reviews. Newark, Del: International Reading Association. 77-117.

Bachman, L. F. and Palmer A. S. (1983) "The construct validity of the FSI oral interview" in J. W. Oller (ed) Issues in Language Testing Research. Rowley, Massachusetts: Newbury House, 154-169.

Bachman, L. F. and Savignon, S. J. (1986) "The evaluation of communicative language proficiency: a critique of the ACTFL Oral Interview." Modern Language Journal 70, 4, 380-390.

Beheydt, L. (1987) "The semantization of vocabulary in foreign language learning." System 15, 1, 55-67.

Berkoff, N. A. (1985) "Testing oral proficiency: a new approach" in Y. P. Lee (ed) New Directions in Language Testing. Pergamon Press, 93-99.

Bialystok, E. (1983) "Some thoughts on the notion of "communication strategy"" in C. Faerch and G. Kasper (eds) Strategies in Interlanguage Communication. London: Longman, 100-118

Blum-Kulka, S. and Levenston, E. (1978) "Universals of lexical simplification." Language Learning 28, 399-115. Revised and reprinted in C. Faerch and G. Kasper (eds) (1983) Strategies in Interlanguage Communication. London: Longman, 119-139.

Butterworth, B. (1975) "Hesitation and semantic planning in speech." Journal of Psycholinguistic Research 4, 75 87.

Brazil, D. (1985) The Communicative Value of Intonation in English. (Discourse analysis monographs no. 8) Birmingham University: English Language Research.

Campbell, D. T. and Fiske, D. W. (1959) "Convergent and discriminant validation by the multitrait-multimethod matrix." Psychological Bulletin 56, 2, 81-105.

Carroll, J. B. (1967) "Foreign language proficiency levels attained by language majors near graduation from college." Foreign Language Annals 1, 131-151.

Carroll, J. B. (1978) "Specifications for an English Language Testing Service" in J. C. Alderson and A. Hughes (eds), 66-110.

Carroll, J. B. (1980) Testing Communicative Performance: An Interim Study. Pergamon Press.

Carroll, J. B. and Hall, J. P. (1985) Make Your Own Language Tests: A Practical Guide to Writing Language Performance Tests. Pergamon Press.

Carter, R and McCarthy, M. (1988) "Lexis and discourse: vocabulary in use" in R. Carter and M. McCarthy (eds) Vocabulary and Language Teaching. London: Longman, 201-220.

Clark, J. L. D. (1980) "Toward a common measure of speaking proficiency" in J. R. Frith (ed) Measuring Spoken Language Proficiency. Georgetown University Press, 15-26.

Cronbach, L. J. (1984) Essentials of Psychological Testing. New York: Harper and Row.

Cruse, D. A (1975) "Hyponomy and Lexical Hierarchies." Archivum Linguisticum 6, 26-31.

Crystal, D. and Davy, D. (1975) Advanced Conversational English London: Longman.

Davies, A. (1978) "Language Testing" Parts 1 and 2, in V. Kinsella (ed) (1982) Surveys 1. Cambridge University Press, 127-159.

Davies, A. (1983) "The Validity of Concurrent Validation" in A. Hughes and D. Porter (eds) Current Developments in Language Testing. London: Academic Press, 141-146.

FSI (1985) Handbook for Language and Culture Instructors. School of Language Studies, Foreign Service Institute.

Fulcher, G. (1987a) "Tests of oral performance: the need for data-based criteria." English Language Teaching Journal 41, 4, 287-291.

Fulcher, G (1987b) "Contextual Hyponomy: a communicative approach to teaching lexis in context." Modern English Teacher 14, 3, 14-17.

Hasan, R. (1984) "Coherence and Cohesive Harmony" in J. Flood (ed) Understanding Reading Comprehension. International Reading Association. 181-219.

Hasan, R (1985) "The Texture of a Text" in M. A. K. Halliday and R Hasan (eds) Language, context, and text: Aspects of language in a social-semiotic perspective. Deakin University Press. 70-96.

Henning, G. (1987) A Guide to Languge Testing: Development - Evaluation - Research. Rowley, Massachusetts: Newbury House.

Hieke, A. E. (1985) "A componential approach to oral fluency evaluation." Modern Language Journal 69, 2, 135-142.

Ingram, D. E. (1984) Introduction to the ASLPR. Australian Government Publishing Service, Canberra

James, P. (1985) "Word Trees." Modern English Teacher 12, 4, 31-34.

Jones, R. L. (1981) "Scoring procedures in oral language proficiency ests" in J. A. S. Read (ed) Directions in Language Testing. (Anthology series 9). RELC. Singapore University Press. 100-107.

Jones, R. L. (1985) "Some basic considerations in testing and proficiency" in Y. P. Lee (ed) New Directions in Language Testing. Pergamon Press. 77-84.

Lado, R. (1978) "Scope and limitations of interview based language testing: are we asking too much of the interview?" in J. L. D. Clark (ed) Direct Testing of Speaking Proficiency: Theory and Application. Princeton N.J.: Educational Testing Service, 113-128.

Lantolf, J. P. and Frawley, W (1985) "Oral Proficiency Testing: A Critical Analysis." Modern Language Journal 69, 4, 337-345.

Liskin-Gasparro, J. (1984) "The ACTFL proficiency guidelines: A historical perspective" in T. V. Higgs (ed) Teaching for Proficiency, the Organizing Principle. Lincolnwood, IL: National Textbook Company. 11-42.

Lowe, P. (1983) "The ILR oral interview: origins, applications, pitfalls, and implications." Die Unterrichtspraxis 16, 230-244.

Lowe, P. (1987) "Interagency Language Roundtable Proficiency Interview" in J. C. Alderson, K. Krahnke and C. W. Stansfield (eds) Reviews of English as a Second Language Proficiency Tests. Washington D.C.: TESOL Publications, 43-47.

McCarthy, M. (1988) "Some vocabulary patterns in conversation" in R Carter and M. McCarthy (eds) Vocabulary and Language Teaching. London: Longman, 181-200.

Morrow, K. (1979) "Communicative language testing: revolution or evolution?" in C. J. Brumfit and K. Johnson (eds) The Communicative Approach to Language Teaching. Oxford University Press, 9-25.

Palmberg, R. (1986) "Vocabulary teaching in the foreign language classroom." Forum 24, 3, 15-20 & 24.

Palmer, A. S. and Groot, P. J. M. (1981) "An Introduction" in A. S. Palmer, P. J. M. Groot and G. A. Trosper (eds) The Construct Validation of Tests of Communicative Competence. Washington D.C.: TESOL Publications, 1-11.

Perrett, G. (1987) "The Language Testing Interview: A Reappraisal." Paper delivered at AILA, Sydney, Australia. Mimeo: Department of Linguistics, University of Sydney.

Raupach, M. (1983) "Analysis and evaluation of communication strategies" in C. Faerch and G. Kasper (eds) Strategies in Interlanguage Communication. London: Longman, 199 209

Sinclair, J. McH. and Coulthard, R M. (1975) Towards an Analysis of Discourse. Oxford University Press.

Sinclair, J. McH (1983) "Planes of Discourse" in Rizvil, S. N. A (ed) The two-fold voice: essays in honour of Ramesh Mohan. Salzburg Studies in English Literature University of Salzburg.

Sinclair, J. McH. and Brazil, D. (1982) Teacher Talk. Oxford University Press

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

Sollenberger, H. E. (1978) "Development and Current Use of the FSI Oral Interview Test" in J. L. D. Clark (ed) Direct Testing of Speaking Proficiency: Theory and Application. Princeton N.J., Educational Testing Service, 1-12.

Spolsky. B. (1973) "What does it mean to know a language? Or, how do you get someone to perform his competence?" in J. W. Oller and J. C. Richards (eds) Focus on the Learner: Pragmatic Perspectives for the Language Teacher. Rowley. Massachusetts: Newbury House, 164-176.

Spolsky. B. (1985) "The limits of authenticity in language testing." Language Testing 2, 1, 31-34.

Spolsky. B. (1986) "A multiple-choice for language testers." Language Testing 3. 2, 147-158.

Stevenson. D. K. (1981) "Beyond faith and face validity: the multitrait-multimethod matrix and the convergent and discriminant validity of oral proficiency tests" in A. S. Palmer. P. J. M. Groot and G. A. Trosper (eds) The Construct Validity of Tests of Communicative Competence. Washington D.C.: TESOL Publications, 37-61.

Stevenson. D. K. (1985) "Pop Validity and Performance Testing" in Y. P. Lee (ed) New Directions in Language Testing. Pergamon Press. 111-118.

Underhill. N. (1987) Testing Spoken Language: A Handbook of Oral Testing Techniques. Oxford University Press.