DOCUMENT RESUME

ED 350 830 FL 020 529

AUTHOR Schrafnagl, Jill; Cameron, Duncan
TITLE Are You Decoding Me? The Assessment of Understanding in Oral Interaction.
PUB DATE 88
NOTE 11p.; In: Grunwell, Pamela, Ed. Applied Linguistics in Society. Papers from the Annual Meeting of the British Association for Applied Linguistics (20th, Nottingham, England, United Kingdom, September 1987. British Studies in Applied Linguistics, 3; see FL 020 520.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Foreign Countries; *Interpersonal Communication; *Language Proficiency; Language Research; *Language Tests; *Listening Comprehension; *Oral Language; *Test Construction; Test Validity

ABSTRACT

A study investigated factors in design of oral language comprehension tests and their relevance in determining actual language proficiency, focusing on the tests' decoding requirements. Task-based, interactive oral language tests were designed: (1) to elicit information about integration of first (L1) and second languages (L2) in the learner's language processing; and (2) to avoid assessment of language skills in isolation. The resulting tests were piloted, and information was gathered on overall task fulfillment and on different aspects of the tasks. One test module, in which the task involved researching an issue and reporting results, illustrates the technique. The examinee's report and interview concerning the research were recorded, and the examinee then wrote a report in L1. The report contains five sections each graded separately on the extent to which the candidate was able to identify and integrate information needed to fulfill the task. Analysis of test performance suggests that conventional oral scales concerned with surface features of language are not adequate predictors of ability to comprehend in oral interactions. In addition, in this test it appeared that comprehension skills were closely related to other complex and integrative skills such as ability to write a coherent report in L1. (MSE)

ARE YOU DECODING ME? THE ASSESSMENT OF UNDERSTANDING IN ORAL INTERACTION

Jill Schrafnagl and Duncan Cameron
Institute of Linguists

This paper addresses the question of the comprehension of spoken discourse and its role in language proficiency and proficiency assessment. It will draw upon various aspects of our work in this area, including data from our recent pilot testing. Our main contention will be that, although comprehension skills are a *sine qua non* in spoken communication, they cannot be assessed *directly* as many language tests, including the US FSI interview and most FL oral tests used in Britain, claim to do. A central concern in our work on L2 proficiency is whether the significance of what is said or written has been recovered by the hearer of reader, or what significance she has imposed on the discourse. We shall be discussing the problems of assessing proficiency in the decoding of spoken language, looking first at how this is currently approached in second language testing. Subsequently, we have a few comments to make on the subjects of language proficiency itself, on proficiency and performance testing, and on comprehension during oral interaction.

Finally, we shall outline our work on the design and development of tests of second language proficiency and in particular their decoding requirements and look at some of the data from pilot tests involving L2 oral interactive performance.

The aim of the project is to develop examinations of foreign or second language proficiency at levels from beginners up to first degree-equivalence for use in the 1990 and beyond. The syllabus should be applicable to all languages likely to be in demand - currently, the Institute of Linguists runs examinations in some 40 languages at lower levels, dropping to 20 at the highest level.

A syllabus for all-comers examinations which set out to test proficiency in using a foreign or second language for communicative purpose, whether the user's ends are social, professional, or both, must be sensitive to the heterogeneity of language learners and users and the multiplicity of their goals. The types of proficiency tested need to be derived not from L2 classroom language use but from how languages are used naturally, by native speakers and bilinguals at any stage in their learning and in personal or professional domains. (It is not possible in this paper to discuss the practice of applying an ideal native speaker model, a monolingual model generally, to the bilingual. We do have considerable reservations about this, and about the convergence implications it has and we have taken a rather different point of departure in our assessment criteria).

The tests we have designed are task-based. The tasks themselves are based in what people need to be able to do when called on to use a second language

in addition to their first, or generally preferred language, in contexts where communication, and not learning, is the primary objective.

The patterns of L2 use which emerge from investigations of so-called real-life language use in context are complex. In performance, a learner's two (or more) languages are integrated in a variety of ways. The second language user also typically lives out his/her role as a bilingual, mediating between cultures and the members of different speech communities. This role as mediator and facilitator of cross-cultural, cross-lingual communication is particularly important at higher levels of proficiency, where learners are increasingly able to, and motivated to, apply their L2 competence not only in personal but also in professional contexts. These features of second language use have hitherto largely been ignored in testing, even in what is declared to be proficiency testing. The conventional approach is to assess L2 proficiency monolingually, the only usual exception being tests of translation or interpreting.

A further tendency is to assess *skills* in isolation, and this too has very little to do with natural language use. The skills and subskills are typically integrated in many different configurations in natural language use, including second language. It only makes sense to de-integrate, or segregate, skills when testing for particular experimental and diagnostic purposes. Where teaching programmes are based on a contrived separation of skills the (achievement) testing that follows will tend to take the same approach; the arguments for this are circular and generally not predicated on any construct of language competence, any theory or descriptive framework for language use or any model of language proficiency.

The terms of our research and development brief require us to design, pilot, modify and retrial proficiency tests and assessment systems in a range of languages and have the new examinations ready to go into production by 1989/90. We have thus had to combine experimental and pilot phases of testing, and are about to enter the third and final phase of piloting. The material we have piloted to date consists of integrated task-based tests involving French, Spanish, Italian and currently Chinese and English. We shall be concentrating in our presentation of data on examples where the oral/aural interactional comprehension skills are an essential component of the language proficiency performance being tested.

First, however, it may be helpful to consider how the comprehension of spoken discourse is currently approached elsewhere in second language testing. It is frequently assumed that the sort of examination task that is conventionally known as "listening comprehension" is a valid indicator of a learner's ability to decode language in its aural form. It is widely recognized, of course, that the assessment of decoding skills is fraught with difficulties; it may still, nevertheless, be worthwhile to look in more detail at some these.

One basic objection is a very familiar one - that the "listening comprehension" exercise has to elicit some sort of observable behaviour in relation to a process which, in natural language use, need not have any visible or observable manifestations. So failure to perform adequately on such tasks may be a result of the nature of the behaviour eliciting task, and not of a deficiency in the aural decoding skill. There are, however, further objections. Aural decoding in a natural context takes place typically within each decoder's mental frameworks; these allow the decoder to predict the drift of a message, to block or ignore irrelevant or uninteresting messages, and select only those bits of message that are relevant for whatever purpose the decoder has at just that moment when he is listening. The skills required by the same decoder when he is participating in a "listening comprehension" exercise without these comfortable props, are rather different; failure to decode a "listening comprehension" exercise cannot, then be taken as evidence of an inability to decode in contexts of natural language use. There is a further serious objection to the interpretation of "listening comprehension" exercises as valid instruments for assessing aural decoding skills. What the classic listening comprehension exercise by its very nature is unable to simulate is that in natural, non-examination, language use, aural messages which are to be decoded are typically negotiable. The listener - or "receiver of the message" - in face-to-face interaction, is able to encourage the supplier of the message to adjust, rephrase or recapitulate the message until he or she is confident of having received and understood just that information that the receiver wants.

(For the purpose of this paper the word *interactive* will be used to describe this sort of listening, realizing that the word has been used in a rather different sense in relation to reading and other listening processes).

The conventional listening comprehension test is a test of non-interactive listening skills. While for many, learners and native speakers alike, an important non-interactive listening scenario is "entertainment" listening - listening to radio, television or plays - non-interactive listening is a more narrowly-based skill in many professional and vocational uses of languages, being restricted to listening to presentations and lectures in a work or learning context. A proficiency examination could not claim that a wide enough range of listening proficiency skills can be assessed solely by means of a non-interactive listening exercise.

It could be claimed that oral examinations are designed to assess interactive listening skills. Unfortunately, many oral examinations are designed typically to elicit a sample of language and then to assess it by means of various fluency/accuracy scales. A different type of oral examination can be devised in which the candidate 'controls' the interaction and negotiates with the interlocutor to elicit a message in a form consonant with his or her ability to decode it.

The relationship between interactants in which one holds information and the other has to apply appropriate interactional strategies to access the information can be simulated in the examination context by tasks in which candidates have to work within an information or opinion collecting context defined by a brief. The conventional interview examination is typically assessed by means of rating scales that concern themselves with such features as quality of language, pronunciation, fluency, and so on - features derived from an analysis of the surface features of language, and not in any way related to the decoding processes that are taking place while the interaction develops.

We need to understand a great deal more than we do know about how people interact to negotiate meaning and the nature of the processing routines involved, about what makes a good interactant and good listener.

In oral L2 testing, such phenomena as pauses, misalignments, signals of comprehension monitoring and confirmation elicitation are too easily dismissed as dysfluencies marking off the L2 speaker's shortfall from the idealized native-speaker behaviour; but they could equally well be efficient and effective interactional strategies employed to get the interlocutor to modify his/her input.

So we need to know more about the way interaction is modified in different situational contexts, particularly, how native speakers and non-native speakers interact. The body of research into interactional competence has focused largely on conversation analysis of L1 speakers, and we should be cautious about generalizing from this to bilingual and L2 contexts.

Comprehension studies also tend to focus on communication breakdown and repair mechanisms. It is not by any means always apparent in oral interactions when or how that breakdown has occurred, particularly not in a test situation.

It must be remembered that we were not concerned with developmental acquisitional issues except very indirectly, in that we are sampling proficiency at different levels. We were seeking to design proficiency tests which would sample performance on tasks simulating major features of those encountered in the use of one's two languages outside the classroom.

The resulting test design at the highest levels, which are very approximately equivalent to A-level and first degree level, is modular. Each module requires testees to perform complex integrated tasks. The modules differ in the area of language use to which L2 proficiency is applied - for example, political and social research, international journalism, media monitoring, negotiation and representation, information services, transactions via the telephone or correspondence. Consequently, the modules differ, too, in the range of texts and tasks which are encountered; they require varying arrays of skills and subskills; and differing roles are assumed by the L2 and L1 in every case. For the examinations themselves, candidates will select from the range of modules those

which best reflect their interests, goals and perceived needs and attempt any module whenever they feel ready. Thus, examinations are not only focused on language use, they are also centred on the user.

In sampling performance on realistic occupationally-oriented tasks, these tests are in some ways similar to language performance testing as developed, for instance, by Marjorie Wesche and her colleagues in the Ontario Test of English as a Second Language. There are also major differences; where OTESL uses banding of skills in a number of separate scales, performance in our tests is assessed entirely on task fulfilment criteria (the "driving test" approach).

During the piloting of tests we have not only been assessing overall task fulfilment but have been collecting data on different aspects of the tasks. The data produced on one particular module form the basis of the following discussion. The module is degree-level, and the task is one of researching an issue using a variety of sources and reporting the results.

In the pilot material from which the data are taken, the L1 was English and the L2 was German. The task specification required the testee to research and report on the situation of the migrant worker, or *Gastarbeiter*, in the Federal Republic of Germany, specifically the policies on assimilation, integration, repatriation and recent changes in approaches to the education of the children of migrant workers. The putative end-user of the report is a British MEP who needs to be briefed for his/her work on a minorities committee. The detailed task remit specified the aspects on which information had to be obtained and evaluated. Two types of source were available: a dossier of printed materials including surveys, editorials, excerpts from government reports and statistics, and an interview with a German representative of a nongovernmental minority rights commission. The interviewee would, if asked, be able to update and flesh out the information from the published materials. The interview subtask thus incorporated an information gap situation with pre-allocated roles. The oral input was contextualised by both the task specification and the dossier of materials. The input information is held constant by also giving the interviewee a detailed specification and by training interlocutors, to ensure consistency of response to the testee's questions. the interviewer - the candidate in other words - has an active planning role, has to produce comprehensible output and manage the interaction so that he/she elicits comprehensible input, and has constantly to check his/her intake and update his/her mental representation. At the same time, the interviewer must record the relevant information, using whatever aids and strategies preferred: note-taking in the L2, extempore translation and notes in the L1, ticking against some checklist prepared along with the interview prompt notes, and so on. Strategies are, of course, not prescribed.

The interviewee will not always stick to the point after all it is the interviewer's point, that is, it is the candidate who is pursuing a particular objective.

So the interviewee may - is in fact briefed to - introduce opinion (flagged as such) along with fact. The candidate has to be able to distinguish reported fact from opinion. The results of the research are then written up in a report in the L1. This is a cognitively demanding task, as are all the other modules and tasks at this level. The assessment of task fulfilment looks solely at the evidence in the report. We are concerned with the different factors in task performance only in so far as they contribute to the required outcome. In our pilot tests, however, we are also investigating relationships between performance on subtasks and overall task fulfilment. We are after all, assuming with our test design that success in the outcome is only attainable if the constituent processes are successful. We are also, less explicitly but no less importantly, claiming that conventional approaches to the assessment of L2 proficiency on separate accuracy and fluency scales for the four *macroskills* are adequate to establish the kind of performance proficiency that will get you and keep you a demanding job using languages. Specifically, the structural-lexical-phonological-fluency scales used in so many tests of second language proficiency reflect a totally inadequate model of interactional language competence. While the examination was being administered the native-speaker German interview interlocutor did not assess the candidates' performance in any way, restricting herself to the role of information supplier as detailed in the task brief. Each oral interaction was recorded on tape-recorder.

Following the interview, in which candidates sought to supplement information gleaned from the written dossier, reports were written in English (that is, the learner's L1) according to the task brief, fleshing out information from different sources. It is important to realise that the carrying out of the task brief depended upon this process of synthesis. It was impossible, in other words, in an adequate report to mark the success of the report according to a check-list of essential points whose origin (from either text or interview) could be clearly identified.

The report contained five sections; each section was awarded a maximum of four points. Each omission or misrepresentation of information in any section led to a subtraction of one point, giving a range in each section from zero to four. This measure on a scale of 0 - 20 (Fig.1 column A) indicates the extent to which each candidate was able to identify and put together the information needed to fulfill the task brief.

Scale B in Fig.1 is concerned more directly with the success or failure of the interview - measuring success or failure as the ability of the candidate/interviewer to extract the necessary information from the interviewee/interlocutor - in other words, to fulfill the terms of the task brief as it applied to the interview

There were seven different topic areas where information from the written sources needed to be complemented by additional or relativizing information

which could be obtained only from the interview. Each report was given a 0 - 7 rating depending on whether the interview information in each of these areas had been accurately and fully transmitted.

It was found that there was a very close association between scales A and B. All candidates who gained 7 point maximum on scale B gained a high score on scale A (candidates 2,4,5,9,11,13,19).

Only one high scorer on scale A gained less than 7 on scale B (candidate 15).

It could be argued that the close relationship between success or failure on these two scales was caused by the overlap between them - both are arrived at by means of an analysis of the same product - the candidates' final report.

Information measured by scale B is also necessary for success on scale A.

Reports were assessed according to a third measure, which would appear, on the face of it, to be concerned with quite different skills from those measured by Scales A and B. Reports were given a score out of nine (see scale C) according simply to their effectiveness as reports, without reference to the information appearing in them. Four points maximum were awarded for the coherence of each report, taking into account logical presentation and use of sources as exemplification. A maximum of 4 points were given for appropriacy of register and clarity of linguistic presentation. One bonus point was awarded for readability. Perhaps surprisingly, success or failure on this largely L1 skill is just as closely related to each of the previous two scales as scales A and B are related to each other. Five of the highest scores on this scale C (2,4,5,9,11) were the highest scorers on each of the other two scales; only one high scorer on scale C (15) was *not* one of the highest in scales B.

A fourth scale - Scale D - was concerned with conventional *quality of language* features in the candidates' spoken language. Marks were awarded by a German native speaker listening to a tape recording of each interaction. The sixteen-point scale consisted of four categories of four points each:

(i) accuracy and range of structure
(ii) accuracy of phonology, stress and intonation
(iii) range and accuracy of lexis
(iv) accuracy

Points from zero to four were awarded impressionistically in each category. High scores on this scale do not have a close relationship with scores on the other scales - for example, candidate 17 got a high score on scale D, but low scores on the other three scales, while candidate 4, with the same score on the oral scale was also one of the high scores in other scales.

Spearman-Brown Rank-Difference correlation coefficients were calculated for all of the cross-correlations between the four scales. The results (Fig.2) bear out the observations made during the preceding discussion. Cross-correlations

between measures A,B, and C are all very high and very comfortably above the 01 significance level for 19 cases. In striking contrast, cross-correlations between measures D - concerned with surface features of L2 language - and the measures derived from L1 report are very low, and do not approach significance level.

It would appear, then, that conventional oral scales concerned with surface features of language are not adequate predictors of ability to comprehend in oral interactions. What is perhaps surprising is that our results indicate that comprehension skills of the type we were concerned with are closely related to other complex and integrative skills - such as the ability to write a coherent report in the L1.

BIBLIOGRAPHY

Duran, Richard P (1984) Some implications of communicative competence research for integrative proficiency testing. In Charlene Rivera (ed): Communicative Competence; Approaches to Language Proficiency Assessment *Multilingual Matters* 9, 44-58.

Hulstijn Jan H (1985) Testing second language proficiency with direct procedures. A comment on Ingram. In K Hyltenstam and M Pienemann (eds): Modelling and Assessing Second Language Acquisition *Multilingual Matters* 18, 147-158.

Ingram, D E (1985) Assessing Proficiency: An overview on some aspects of testing. In K Hyltenstam and M Pienemann (eds): Modelling and Assessing Second Language Acquisition *Multilingual Matters* 18, 215-276

Seaton, Ian (1987) Comments on Wesche: Second Language performance testing: The Ontario Test of ESL as an example. *Language Testing* 4,1: 48-54.

Spolsky, Bernard (1986) A multiple choice for language tests. *Language Testing* 3,2: 147-158.

Wesche, Marjorie B (1987) Second Language performance testing: The Ontario Test of ESL as an example. *Language Testing* 4,1.

FIG 1

	A REP INF		B INT TOP		C REF FEA		D CONV OR SCL	
	Score	RO	Score	RO	Score	RO	Score	RO
1	5	16	4	10	4	15	9	12
2	17	1	7	1	1	2	9	12
3	4	17	2	19	1	19	8	15
4	14	7	1	1	7	2	13	2
5	15	6	7	1	7	2	8	15
6	11	11	4	10	5	10	13	2
7	9	13	4	10	4	15	6	19
8	8	14	3	18	4	15	8	15
9	16	4	7	1	7	2	9	12
10	10	12	4	10	5	10	11	7
11	17	1	7	1	7	?	10	10
12	12	10	4	10	5	10	13	2
13	17	1	7	1	8	1	10	10
14	13	8	4	10	5	10	14	1
15	16	4	5	8	7	2	11	7
16	13	8	5	8	6	3	13	2
17	5	16	4	10	5	10	13	2
18	6	15	4	10	4	15	8	15
19	13	8	7	1	6	3	11	7
MAX	20		7		9		16	

Score and Rank order data on four assessement scales

RO Rank Order
A REF INF Report Information
B INT TOP : Interview Topic Recovery
C REF FEA : Report Writing Features
D CONV OR SCL : Conventional Oral Accuracy and Fluency Scales

FIG 2

	A	B	C
B	.875*		
C	.937*	.933*	
D	.181	.176	.269

N = 19
.665 = Sig. .01

A Report Information
B Interview Topic Recovery
C Report Writing Features
D Conventional Oral Accuracy and Fluency Scales