DOCUMENT RESUME

ED 266 639 FL 015 453

AUTHOR Schreuder, Klaas
TITLE Formative Tests of Speaking Proficiency in the Mothertongue for 14-Year Olds.
PUB DATE 85
NOTE 13p.; In: Practice and Problems in Language Testing 8. Papers presented at the International Language Testing Symposium of the Interuniversitare Sprachtestgruppe (IUS) (8th, Tampere, Finland, November 17-18, 1984); see FL 015 442.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Adolescents; *Dutch; Foreign Countries; *Language Proficiency; *Language Tests; Native Speakers; Secondary Education; Second Language Learning; *Speech Communication; Teacher Attitudes; *Test Construction; *Test Reliability

ABSTRACT

The development and validation of oral native language tests for Dutch adolescents was conducted for both assessment and curriculum development purposes at the secondary level. The oral language situations tested were the monologue, dialogue, and polylogue or group conversation on a topic. An American inventory of common speech subjects was used to construct test situations, and the test subskills were drawn from the students' textbooks. The situations, subjects, and subskills were combined into three tests, which were administered to 14-year-olds in all areas of the country and all dialect regions. The tests were taped, and teachers were asked to rate the dialogue and monologue tests on content, organization, language, delivery, and communication. The interrater reliability and homogeneity of assessment criteria were analyzed statistically. On the whole, it was found that the two test types were closely correlated and that overall scores would have sufficed for ranking students in the case of almost all raters. However, when only one rater is available, as in a classroom testing situation, the analytic method of scoring is preferred to the overall scoring method. (MSE)

FORMATIVE TESTS OF SPEAKING PROFICIENCY IN THE MOTHERTONGUE FOR 14-YEAR OLDS

Klaas Schreuder (Cito, Netherlands)

1 Introduction

For oral tests to be used in the course of the education process and meant to support it, that is for formative oral tests, as a rule only direct and not indirect testing methods will be adopted. That is to say, in taking such an oral test pupils will actually have to speak. And, indeed, I think that indirect measures of the oral skill at an intermediate level should be avoided, not only or in the first place for reasons of validity but with a view to

- the motivation of pupils, as well as
- the effects such tests have on education, whether intended or not.

Generally speaking testconstructors prefer their tests to elicit behaviour that does not differ too much from the criterion behaviour, the sort of behaviour that would occur in ordinary life and which is, in fact, what the test is meant to judge. In everyday reality there are, generally speaking, three situations calling for a demonstration of the oral skill: the monologue, the dialogue and the polylogue. In the large majority of situations falling into any of these three main categories, each speaker can, to a large extent, determine the direction of the activity in which he is engaged. This is selfevident in case of the monologue; in the other cases the influence exerted by the participants will not exactly be alike, but however that may be, it cannot be predicted with any certainty what the next speaker is going to say. Even situations that limit a speaker's free scope severely such as interviews, e.g. between a senior and a junior staff member, even such situations remain open-ended to a large extent. The junior staff member can cause the event in which he participates to take an unexpected turn. If it is thought important that the behaviour elicited by the test and the criterion behaviour do not diverge too much, the test should create a situation which leaves the testee the scope he would have in everyday reality.

Apart from this, 'realistic' tests will give rise to fewer methodological problems in validating tests, that is, provided the test's reliability meets the proper requirements. I take for granted that a test which is reliable and prompts behaviour that is quite close to criterion behaviour can be considered a valid test. Only if the behaviour required by a test deviates more markedly from the behaviour that is the real object of the evaluation, further validation is called for.

Of course I should make mention at this point of a difference between testing the oral skills of L1 and L2 learners. L1 learners can be assumed to avail themselves of a certain latitude in giving

direction to a conversation. But L2 learners at an elementary level will not regret the constraints put upon their freedom of expression: they would not know to use it if it were granted to them. That is why strictly structured tests of simple vocabulary, pronunciation or sentence structure are not at variance with the general principle of allowing the testee free scope. Incompatibility only arises at an advanced level of L2 proficiency. At this level requirements will be made of the L2 learner that come close to those made of the L1 learner. What I am going to say about oral tests of Dutch for 14 year old Dutch pupils might consequently be of interest to those concerned with testing L2 proficiency.

First of all I shall say one or two things about the starting points adopted for the construction of the tests. Then I shall discuss the test formats and present the results of the pretests and rating try-outs.

2 Test content

The oral tests for native speakers of Dutch developed so far are part of a plan to compose a set of oral tests for the whole of secondary education. Tests for intermediate levels have been constructed first, because it is supposed that a series of such tests would contribute to the development of programmed instruction in the oral skill. That is to say, individual teachers as well as writers of educational material could benefit from the existence of a set of formative tests based on relevant educational objectives. An inventory of educational objectives through an analysis of most frequently used textbooks showed a lack of any sort of systematic approach.

The series of tests was set up as follows.
1 Oral proficiency was defined as: the skill to express oneself orally in an adequate manner in functional language situations. The latter are considered to be language situations derived from an analysis of requirements made of pupils both at school and out of school. The main criterion for categorizing these situations is the number of speakers involved: monologue, dialogue and polylogue. Each of the three main divisions is to comprise about ten language situations in accordance with the subcategories coming under the headings of monologue, dialogue and polylogue.
2 In selecting relevant subjects and situations we used the inventory presented in 'Basics in speaking and listening for High School Graduates' by Ronald E. Bassett et al. in Communication Education (1978: 293 f), which is an outcome of the minimal competencies movement in the USA. It lists all the situations calling for oral speech in which the average (American) citizen may get involved, situations that have to do with one's occupation, with citizenship and maintenance (private life). With a few adaptations to the Dutch situation, the list proved very helpful.
3 The tests' contents were largely determined by the relevance of subskills as apparent from their featuring in textbooks. Such subskills are:

reformulating, reasoning, distinguishing major matters from minor ones, ranking facts chronologically, expressing an opinion and (for the categories other than the monologue) asking and answering questions, contradicting, responding to the expression of emotions. 4 The way in which we wanted assessments to be made, was based on the following consideration. Since the tests are meant for classroom use enabling teachers to ascertain whether or not their pupils command a particular component of oral proficiency after instructions or without special training as the case may be-analytic assessment on the basis of a number of criteria will be more in the interest of the pupil and the teacher than global assessment. In this case agreement among raters would seem to be less important than consistency of the ratings over a period of time.

I shall now explain briefly with the help of some examples how the elements of this test content have been combined into tests. These elements are:

- speech situation (number of speakers)
- subject matter/context
- subskill.

A ('monologue') test puts the pupil in a situation in which he has to report to others on something that he has seen or heard. In this case it goes like this: a pupil is shown a ten minute film about the bio-industry. After some minutes' preparation he then has to report on the film to his class. How he goes about this is up to him. To aid him he does get a list of a number of obligatory points. Such a list proved an essential aid to memory for a great many pupils. The report is expected to be four minutes long. It is clear what we are after in this assignment: the pupil has to give a report (a subcategory of the monologue) in a situation which might occur at school or in a future job and in doing so he has to demonstrate the subskill of distinguishing between major and minor matters.

A second ('dialogue') test puts the pupil in a situation in which he has to carry on a conversation, not the informal everyday chat, but rather a more or less formal interview. The pupil is expected to acquire information by asking questions in some everyday situations. For instance, he is supposed to be wanting information about a periodical that he might subscribe to, about a sporting club that he is thinking of joining, about a particular organized trip that he might want to enter for. The pupil receives a sketch of the situation with the assignment to ask, within a five minute span, all the questions necessary to obtain the required information. Preparation time is ten minutes. After this there is the interview during which it is up to the teacher to supply the information asked for, but not all at once: on a number of occasions which the teacher selects himself he gives only evasive answers requiring the pupil to keep on asking questions until he has received a complete and clear answer. The pupil must not let himself be put off.

The subject of a third test, a 'polylogue', comes under 'citizenship', somewhat awkwardly perhaps, seeing that the test is meant for 13 to 14 year olds. Such a subject is 'public transport'.

Four pupils are given texts about public versus private means of transport, some days before the test is taken. Two of the pupils get a text which is biased in favour of private transport, the text the two other pupils get supports public transport. After studying their texts these four pupils have to carry out a particular assignment in discussing the subject. In principle the contributions of all four pupils are rated.

By means of these test formats we try to imitate as large a sample from real-life situations in which oral speech is used as possible.

3 Pre-tests and rating sessions

One of the constraints involved in the test situation is implied in the question whether the behaviour that is elicited can be assessed properly. To answer that question pretests and rating sessions were organized, the results of which I will discuss now. By means of the pretests we hoped to find out what was the best way of assessing the pupils' achievements. In the pretests the following tests were used: the monologue-test Report on a Film; the dialogue-test Asking Questions; the polylogue-test Discussion PUBLIC TRANSPORT.

The tests were set to pupils about 14 years old in second forms of secondary modern and lower vocational education. Each test was done by fifty pupils from various dialect areas all over the Netherlands. With a view to the rating sessions that would take place at a later date, the test sessions were taped: the monologue- and dialogue-tests on audiotape, the group discussions on video. Four rating sessions have taken place. The raters were teachers of Dutch who had some experience with the second form. Three of the four sessions will be discussed here.

These sessions were set up like this.

REPORTED RATING SESSIONS

No.	NUMBER OF RATERS	TEST	RATED CAND.	CRITERIA	SCALE POINTS	GLOBAL RATING
1.	11*	DIALOGUE* ANSWERING QUESTIONS	21*	11*	4*	X
2.	11*	DIALOGUE* ANSWERING QUESTIONS	21*	11*	4*	X
3.	12	MONOLOGUE REPORT OF FILM	27	4	4	X

*IDENTICAL

At the first two sessions, separated by a week's time, the performances of 21 pupils on the dialogue test were assessed. At the third session the achievements of 27 pupils on the monologue test were evaluated. At all sessions analytic assessments on the basis of a set of criteria were made. Raters were also asked to give a global judgement on the basis of the usual Dutch ten-point scale. The following rating formats were used: an extensive model featuring eleven criteria and a four-point scale at the first two sessions, a more concise model with four criteria and a different four-point scale at the third session. The formats look like this.

RATING MODEL SESSION 1 and 2 DIALOGUE

CONTENT	SCALE POINTS YES	NO
1. brings up the obligatory points	o	o
2. brings up points of his own choosing	o	o

	RARELY	SOME TIMES	OFTEN	ALMOST ALWAYS
3. asks for necessary elucidation	o	o	o	o
4. repeats himself unnecessarily	o	o	o	o
LANGUAGE				
5. words and sentences are suitable	o	o	o	o
6. pronunciation and words are non-standard	o	o	o	o
DELIVERY				
7. speaks clearly	o	o	o	o
8. speaks fluently	o	o	o	o
9. speaks monotonously	o	o	o	o
COMMUNICATION				
10. maintains contact with interlocutor	o	o	o	o
11. interrupts the interlocutor	o	o	o	o

GLOBAL RATING (marks from 1-10)

RATING MODEL SESSION 3 MONOLOGUE

CRITERIA	SCALE POINTS WEAK	BARELY SUFFICIENT	FAIR	EXCELLENT
1. CONTENT	0	0	0	0
2. ORGANISATION	0	0	0	0
3. LANGUAGE	0	0	0	0
4. DELIVERY	0	0	0	0

GLOBAL RATING

3 1 Consistency and reliability of raters

The first question I mentioned above concerned the reliability and consistency of the raters. Consistency is, of course, an aspect of reliability. The consistency of raters could be computed as there were two sessions, a week apart, at which the same taped performances were assessed. We were particularly interested in this aspect because of the future use that would be made of the tests: as teachers will use them in class and as they will not be able to enlist the help of other raters, it is much more important for teachers to be consistent in their assessments than for teachers to agree with each other: consistency is more important than inter-rater agreement.

TABLE 1. COEFFICIENTS OF CONSISTENCY IN DIFFERENT RATING SESSIONS

RATER 1	.86
RATER 2	.85
RATER 3	.79
RATER 4	.84
RATER 5	.73
RATER 6	.73
RATER 7	.80
RATER 8	.73
RATER 9	.79
RATER 10	.74
RATER 11	.86

These correlations are relatively high which means that the raters are fairly consistent in their judgements.

In order to determine the rater's reliability - in the sense of their homogeneity - a homogeneity analysis was used. This procedure assigns numbers to scalepoints for each rater. These numbers are called scalevalues. By means of these scalevalues the original data (the rating sheets filled out by the raters) can be translated into a numerical datatable. For each rater the correlations between the perceived rating and the mean (= true) rating can be computed. The square of that correlation is the rater's reliability.
(Homogeneity analysis determines the scalevalues for the scalepoints in such a way, that the mean rater reliability - or homogeneity - is as high as possible.)
The rater's reliabilities for each of the three sessions were as follows.

TABLE 2. RATER RELIABILITIES-DISCRIMINATION MEASURES

RATER	RATING SESSION 1	2	RATER	RATING SESSION 3
1	.80	.81	1	.43
2	.74	.74	2	.67
3	.62	.69	3	.38
4	.71	.73	4	.58
5	.55	.59	5	.61
6	.50	.72	6	.43
7	.68	.67	7	.55
8	.63	.64	8	.43
9	.84	.76	9	.65
10	.63	.59	10	.51
11	.73	.75	11	.70
			12	.62
Mean	.68	.70		.55

The difference in mean reliability between the first two sessions and the third can perhaps be accounted for by the fact that different rating formats were used: the first format used eleven criteria and four scalepoints, the second format used four

criteria (and again four, but different scalepoints). The four criteria of the second format may be too complex to be applied unambiguously.
High mean reliability does not mean of course, that all raters agreed in their assessments. The analysis only determines the raters' homogeneity, that is, the extent to which their assessments pair, or also the extent to which one rater's assessment can be predicted on the basis of another rater's assessment. The raters will only give more or less similar assessments, if the scale values that are assigned to the scale points for each rater resemble each other per scale point, that is, if they have only little variance. The scale values obtained in the three sessions can be summarised as follows.

TABLE 3. SCALE VALUES. MEANS AND STANDARDDEVIATIONS.

RATING SESSION	SCALEPOINT	MEAN SCALE VALUE	STANDARD DEVIATION
1	RARELY	-.77	.07
	SOMETIMES	.00	.19
	OFTEN	.96	.19
	ALMOST ALWAYS	1.32	.18
2	RARELY	-.84	.09
	SOMETIMES	.11	.18
	OFTEN	.97	.16
	ALMOST ALWAYS	1.24	.19
3	WEAK	-1.21	.22
	BARELY SUFFI.	-.13	.14
	FAIR	.62	.11
	EXCELLENT	1.32	.26

From the differences between the mean values it can be concluded that the raters frequently used the same scalepoints, which means that they agreed in their assessments. For if this were not the case, if they had differed more or less randomly time and again, the mean scalevalues would have been identical.

From the fact that, in all three sessions, there is a regular increase of the mean value (even though the scale points used are not the same) it is clear that the scale points have been interpreted similarly to a certain extent, which is to say that, for example,

'often' has not systematically been regarded as superior to 'sometimes' in case of a particular criterion by some raters, whereas other raters dealt with this scale point the other way round. The differences between the mean values clearly show when reliability intervals are computed on the basis of the standard deviations found. As far as the column of standard deviation is concerned: the smaller the SD of the scalevalues assigned to a scalepoint is, the more frequently this scalepoint has been used by all raters at the same time. These standard deviations are low. Those for 'weak' and 'excellent' are highest for the third session. This is in accordance with the relatively low mean rater reliability for this session, as was shown in TABLE 2: these two scalepoints were apparently interpreted differently by different raters.

3.2 Independence of criteria

The second question regarded the independence of criteria. It was expected that the criteria could not be applied entirely independently, since they relate to one complex skill, but on the other hand they were not supposed to be connected too closely, as in that case far fewer criteria or even one criterion would suffice. And then the so-called analytic assessment would not have any advantages over a global one. Analytic assessment has the advantage, with classroom use of tests, that the pupils and their teacher can learn from the results at what points the pupils' skills fall short and call for additional training.

The homogeneity of the criteria has been determined by means of the same scale analysis was was used for computing rater reliability. For each of the three sessions a mean criterion was determined. If all criteria correlated high with this mean criterion, in fact one criterion would suffice. The next table shows the square correlations between the criteria and the mean criterion.

TABLE 4. DISCRIMINATION MEASURES CRITERIA

CRITERIA	RATING SESSION 1	2	CRITERIA	RATING SESSION 3
1	.07	.00	1	.62
2	.01	.01	2	.72
3	.25	.18	3	.59
4	.03	.07	4	.66
5	.68	.70	MEAN	.65
6	.41	.48		
7	.57	.67		
8	.57	.55		
9	.40	.57		
10	.57	.40		
11	.02	.00		
MEAN	.33	.33		

The mean correlations are low, particulary for the first two sessions. This means that the critoria do not overlap completely. This is mainly due to the criteria in the first category, that of CONTENT. Together with criterion no. 11 - 'interrupts interlocutor' - they are deviant. The four criteria used in the third session, which are as a matter of fact the headings of the four categories in the first format, all correlate more closely with the mean criterion. (These correlations cannot be simply compared to the others, because the criteria as well as the scale points differed from those used in the first two sessions.)

In order to establish to what extent the categories themselves are homogeneous, the correlations between the various criteria and the mean criterion per category have been computed as well. The next table shows a survey of the (square) correlations.

TABLE 5. DISCRIMINATION MEASURES PER CATEGORY OF CRITERIA

		SESSION 1	SESSION 2
CONTENT	1	.32	.00
	2	.26	.05
	3	.47	.66
	4	.24	.55
LANGUAGE	5	.78	.78
	6	.78	.73
DELIVERY	7	.71	.72
	8	.58	.69
	9	.58	.70
COMMUNICATION	10	.56	.57
	11	.66	.67

Once again, the criteria in the category of CONTENT appear to be deviant: they ill represent the category into which they fall.

The intercorrelations between the categories themselves can be clearly seen from the correlations between the mean criteria per category in the next table.

TABLE 6. CORRELATIONS BETWEEN CATEGORIES

RATING SESSION 1			RATING SESSION 2		
cat 2	.25		cat 2	.17	
cat 3	.27	.60	cat 3	.22	.66
cat 4	.05	-.01 .06	cat 4	.13	.04 -.19
	cat 1	cat 2 cat 3		cat 1	cat 2 cat 3

Categories 2 LANGUAGE and 3 DELIVERY appear to correlate most closely. (Note that the minus signs are due to the peculiarities of the method for scale analysis used and can be ignored.)

From these data it appears that
- the CONTENT category should be split up;
- the criteria of the other categories represent the categories into which they fall fairly well;

- reducing the number of criteria to no more than the category headings is unnecessary,
- the categories of LANGUAGE and DELIVERY in the first format are the ones that could be combined

3.3 Analytic and global rating

The last question that we asked ourselves regarded the qualities of the two ways of assessment: global and analytic. The collected data do not point to a definitive conclusion. We preferred the analytic assessment, and this preference is not ruled out by the results. Here are some figures

1 The correlation between the global marks obtained in the first two sessions is .69.
2 The correlations between the scores on the mean criterion and the global marks given by the raters are
session 1: .78
session 2: .55
session 3: .89
3 The first principal components of session 1 and session 2 (as far as the analytic assessment is concerned) show a correlation of .71. This value is to be seen as a test-retest reliability. From this it appears that
- there is a close correlation between the two methods of assessment,
- global marks would have sufficed for a ranking of pupils in the case of 11 raters,
- for classroom use, when only one teacher can assess the achievements, the analytic assessment is to be preferred.

4 Bibliography

Bassett, R.E. et al. 1978. 'Basics in speaking and listening'. Communication Education 1978: 293 f.

Nishisato, S. 1980. Analysis of categorical data, dual scaling and its applications.
Toronto University of Toronto Press

Rijlaarsdam, G.C.W. 1982. Beoordelen van discussievaardigheid. Amsterdam SCO.