DOCUMENT RESUME

ED 343 401 FL 020 085

AUTHOR	Weeren, Jan van
TITLE	Testing Pronunciation.
PUB DATE	87
NOTE	10p.; In: Equality in Language Learning. Proceedings of the Nordic Conference of Applied Linguistics (5th, Jyvaskyla, Finland, June 4-7, 1987); see FL 020 065.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Applied Linguistics; Comparative Analysis; *Error Analysis (Language); *Error Patterns; French; German; *Language Tests; *Pronunciation; Secondary Education; Secondary School Students; *Second Languages; *Test Reliability; Test Validity

ABSTRACT

Pronunciation is an important subskill in second language learning, therefore worth evaluating. Its quality is commonly assessed in a global, impressionistic way by having learners read aloud. While this allows comparison of examinees' skills, ability to read aloud is a possible confounding variable. An alternative method is to have learners read texts in which only marked elements are judged as correct or incorrect. A study compared the reliability and validity of this and the traditional, holistic method. Pronunciation tests were developed for Dutch learners of French and German that incorporated words in which pronunciation errors occur frequently. The tests were administered to secondary school students (German=26, French=19). Recorded readings were rated in both the traditional way and with the marked-item method by teacher panels to measure reliability. Results suggest the new method can improve evaluation. To determine validity, another experiment simulated regular teacher evaluation of student pronunciation in French by comparing ratings of pronunciation in: (1) spontaneous speech; (2) a traditional read-aloud text; and (3) marked items in a read-aloud text. These results suggest that (3) marked items in a read-aloud text. These results suggest the new alternative is not preferable to traditional, holistic rating because it can not give a general impression of pronunciation quality. Further research is recommended. (MSE)

TESTING PRONUNCIATION

JAN van WEEREN
National Institute for Educational Measurement (Cito)
The Netherlands

ABSTRACT

Pronunciation is regarded as a valuable subskill in foreign language teaching and testing. Its quality is commonly assessed in a global way by having testees read aloud. A more systematical and transparent approach to evaluation is proposed. The reliability of the experimental approach is compared with the reliability of a traditional rating procedure. In a second study the external validity of the experimental approach is determined.

1. Introduction

There are good reasons for not neglecting pronunciation as an important subskill in foreign language teaching and testing. First of all, with a good pronunciation one can make oneself better understood. It gives one's oral production a certain redundancy and this can help a speaker to get his message across more easily. Secondly, pronunciation is quite attractive as a learning objective because of its high pay-off. The number of sounds, sound clusters and intonation patterns in a language is finite, just like the alphabet. Once the system has been mastered it can be applied in one's future performance, thus giving it fundamentally an infinite scope.

Thirdly, a deviant pronunciation means that one is immediately 'marked' abroad. A 'foreign accent' is often used as the means to distinguish non-natives from native speakers. There can be some discussion about the proposal that foreigners should cherish their rôle as non-natives. However, one can think of learners who would prefer a native-like pronunciation, for example those with a particularly strong integrative motivation, and one can imagine situations in which one wishes to avoid immediate identification as a foreigner.

If pronunciation is worth teaching, it is worth evaluating. For this evaluation one needs a speech sample of the testee. This can be obtained

by involving him in a more or less spontaneous conversation or by asking him questions about a text he has read before, about his hobbies, his plans for the future, etc. However, it is difficult to separate pronunciation from other aspects of the testee's oral production, such as fluency, vocabulary, and grammatical accuracy, and judge it independently (Clark and Swinton 1980).

If the testing focus is on pronunciation, it is a common practice to have the testee read aloud a few lines or paragraphs of a written text. The advantage of this procedure is that one can obtain comparable speech samples from different testees. It cannot be denied, however, that this elicitation procedure evokes an unknown variable: the ability to read aloud. The impact of this variable is unclear. As far as we know, no relevant research has been carried out in order to determine its effect. It could cause severe problems with testees who are to some extent dyslectic.

Nevertheless, reading aloud is very common in language teaching and usually testees are familiar with it. Moreover, we cannot neglect the advantage of obtaining comparable speech samples. Having them collected is necessary in order to determine the quality of testing procedures.

Normally, rating is done by giving the testee a mark for the overall impression of his pronunciation. Rating can be based on a four-, five- or six-point-scale. Every point on the scale is defined by a description of one or two prominent characteristics of the testee's performance, like 'occasional phonemic errors, but generally comprehensible' or 'many phonemic errors: very difficult to perceive meaning' (Clark 1972:93). The selection of a certain mark or scale point can be both supported and motivated by a written record made during the testee's performance in which pronunciation errors are successively noted down. The accountability and transparency of such a rating procedure can be improved by working out such records in a systematical way and by letting the final mark depend on it entirely. Furthermore, a systematic and complete record can be used for diagnostic purposes. Weak points in a testee's pronunciation can be subjected to subsequent training. Proposals for setting up such records are found in several testing guidelines (eg. Harris 1969:86-87, Heaton 1979:84-86, and Madsen 1983:66-68). These procedures have in common that in a running text or in a list of unconnected sentences possible pronunciation problems, including stress and intonation problems, are marked; for example (Harris 1969:86).

While Mr. Brown read his newspaper,	primary stress
his wife finished packing his clothes	voiced final consonant(s)

for the trip. The suitcase was already quite full, and she was having a great deal of difficulty finding

vowel quality
primary stress

An unmarked version of the test is read aloud by the testee and the tester ticks off the places where an error is made.

2. Statement of the problem

We decided to examine the reliability of a rating procedure for the evaluation of pronunciation by which only marked elements in a text read aloud are judged in a rather straightforward way: the realization of each element could be judged either as correct or as incorrect. The reliability of this procedure was to be compared with the reliability of the common practice when an overall impression mark is given for the quality of the testee's pronunciation after reading a text aloud. We will call the first, the experimental method, 'atomistic' and the second, the traditional method, 'holistic'.

2.1. The Pretest

First of all, an inventory was made of common pronunciation errors in German and French as spoken by Dutch people. Such inventories can be extracted from several textbooks for the teaching of the pronunciation of German and French. Two texts were selected and adapted slightly. On an average, every 5th or 6th word (German) and every 8th or 9th word (French) contained a pronunciation problem. The German text was read aloud by 10 adults, the French text by 14. All testees had learned the respective foreign language in the secondary school. The testees' tape recordings were marked by a panel of fully qualified teachers, 7 for both German and French, after a short instruction. The tape recordings were presented only once, without stops or replays. The raters were to mark the separate items on a rating sheet of the type that was outlined above. Only incorrect realizations of elements in the texts were marked, so that for each testee and rater the number of pronunciation errors could be calculated

However, unmarked items are ambiguous with regard to their interpretation. An unmarked item can mean two things:

(1) no pronunciation error has been made; or
(2) a pronunciation error has not been registered.

Items that are hardly marked by the raters at all are of no interest. The testees' pronunciation is either correct or errors are not evident enough. Such items are superfluous if the testees are to be discriminated on the basis of the quality of their pronunciation. Items that are marked as incorrect to a substantial extent are more important in this respect.

For a good item, ie. an item with a certain discriminative power, registration of errors is a necessary, but not a sufficient, condition. It is also necessary that a vast majority of raters should react in the same way on the same item and the same testee. Ideally, every rater should mark an item as incorrect each time that it is realized incorrectly and do nothing if it is realized in a correct way. In that case the procedure is absolutely reliable. However, as a rule absolute reliability cannot be attained. Factors such as lack of power of perception or flagging concentration among the raters will prevent this. Nevertheless a certain minimal requirement of reliability must be met if the rating procedure is to be worthwhile.

In order to select successful items for the testing procedure under investigation we set the following conditions:

(1) the majority of raters (5 out of 7) should register an error made by a certain testee on a specific item;
(2) the raters should show a certain measure of agreement in their ratings of all the testees on a particular item. A minimum value of .70 for an interrater reliability coefficient was set.

(As an index for agreement we used percentage of agreement uncorrected for chance.) About 30% of the German and French items met both criteria.

There are reasons to believe that there is some systematic difference between successful and unsuccessful items. In general, successful items do not allow grades of correctness in their realization. They are either realized correctly of incorrectly but not 'more or less' correctly. For example, the realisation of is either correct or incorrect, whereas a vowel can be more or less diphthongized or nasalized. Here tolerance can vary among raters. In the same way the aspiration of a unvoiced plosive can be more or less pronounced. There is no clear cut-off point.

Both tests were revised; the main difference between the resulting German and the French test was that the first was revised in a strictly mechanical way: only pronunciation problems represented by items that passed the pretest were selected, whereas the French test was adjusted

with the help of the actual tape recordings. Almost all the original pronunciation problems were preserved in the definitive version.

2.2. Test and analysis of data

For the trial of the definitive tests new tape recordings were made. The text was read aloud by pupils in secondary schools who had learned German for 2-4 years and French for 3-5 years.

There were 28 testees and 24 items for German and 19 testees and 40 items for French. The tape recordings were rated by 14 fully qualified teachers of German and 15 fully qualified teachers of French, again after a short instruction. The rating procedure was carried out in the same way as in the pretest.

The purpose of the data analysis was to estimate the reliability of this proposed atomistic procedure. Reliability can be expressed in a coefficient between 1 and 0; 1 means that the test measures the differences between the testees in a very systematical, perfect way: there is complete agreement among raters. Every item puts the testees in the same rank order from higher to lower ability. 0 means that the test is completely unreliable, that there is no systematical variance at all.

The following reliability coefficients were found, using generalizability theory (Bolus et al. 1982, Brennan 1983).

	French	German
1 rater/	.86	.52
2 raters	.92	.62

2.3. Setting a criterion

So far we have only compared the results of the French and German experimental tests. It would be helpful to know if the new procedure would entail an improvement of traditional practice. In order to answer this question the tape recordings were played once more and the raters were asked to give the testees a traditional impression mark on a twenty-point scale (that is: the ten-point scale with half points as is commonly used in

the Netherlands). This time, the following generalisability coefficients were found:

	French	German
1 rater/	.79	.62
2 raters	.88	.76

A comparison of these figures with those of the atomistic tests shows that the French test is superior to holistic rating and that the German test is not.

2.4. Preliminary conclusions and discussion

There seems to be a chance that traditional practice in the testing of pronunciation can be improved if an atomistic test is used. Such a test can further the transparency and accountability of the evaluational procedure. It should consist of about 40 items and be designed on the basis of recorded performances. With an inventory of pronunciation problems as a starting point one should try to find regular pitfalls in a text as it is spoken.

In this study we focussed on the reliability of a rating procedure that tried to make the evaluation of pronunciation more explicit and transparent.

It should be stressed, however, that there could be factors involved in a holistic rating that are not covered by the atomistic testing procedure, such as prosody, liveliness, and overall articulation. Therefore we decided to carry out further research into the validity of the procedure. And, last but not least: the validity of evaluating pronunciation on the basis of a text that is read aloud has not been determined yet.

3. Validity study

In order to determine the validity of the atomistic procedure, a new experiment has been set up for French. In this experiment we tried to simulate the regular procedure of a teacher evaluating the pronunciation of his or her own pupils. One can conceive of three different ways of judging pronunciation in such a 'natural' setting:

(1) by holistic rating of spontaneous speech;
(2) by holistic rating of a read aloud text; or
(3) by atomistic scoring of a read aloud text.

elicitation procedure / rating procedure	spontaneous speech	reading aloud
holistic	1	2
atomistic		3

The rating of spontaneous speech is considered most valid for there is no interference with a testee's ability to read aloud. If there is a fairly strong positive correlation between the outcomes of an atomistic scoring of reading aloud and a holistic rating of spontaneous speech, then the experimental method should be considered valid as well. If the correlation between holistic rating of spontaneous speech, on the one hand, and holistic rating of reading aloud, on the other, is sufficiently high, then the assumption that reading aloud and spontaneous speech will yield a different quality of pronunciation is falsified, at least for our target group, pupils in secondary education one year before their final examination.

3.1. Design

The experiment was set up in such a way that 5 teachers of French each had to evaluate the pronunciation of 10 of their own pupils. The pupils' voices were tape-recorded. First, the teachers listened to 10 testees reading a text aloud. They had to note down a global impression mark for each testee's quality of pronunciation. Then, they had to judge the same testees reading the same text according to the experimental atomistic method. Next they heard samples of French spoken spontaneously by the testees. They were, however, put in a different sequence in order to reduce a possible recall effect.

Samples of spontaneous speech were elicited by rôle-playing. The testee was involved in a situational dialogue with a real native speaker.

playing the part of a foreign acquaintance who could provide a holiday job or a police officer who had to cope with the problems of a stolen bicycle.

Again they were asked to note down a global impression mark. In order to assess the reliability of the concurrent procedures the complete cycle had to be repeated with the same teachers and the same testees but with another text and another sample of spontaneous speech. This is a rather strong form of reliability assessment. It is estimated by comparing parallel testforms. The method was not restricted to a simple test-retest procedure with the same test.

The following reliability coefficients were found:

1	.73 < .84 < .91	
2	.65 < .79 < .88	
3	.75	(KR 20)

From these results it follows that the three test formats are approximately equivalent if we take the usual 95% confidence limits into account.

Holistic rating as well as atomistic scoring are proved to be quite reliable. Now, if there is a fairly strong positive correlation between both procedures, then atomistic scoring and holistic rating measure the same variable. The following correlation coefficients were found (again with 95% confidence limits):

	2	3
1	.72 < .84 < .90	.27 < .51 < .70
2		.36 < .58 < .74

The correlations between holistic rating and atomistic scoring are only moderate, whereas the correlation between holistic rating of spontaneous speech and reading aloud is quite high.

3.2. Conclusion

On the basis of this validity study the following conclusions can be drawn. In the upper forms of secondary school one can give value judgements for pronunciation by having the testees read aloud a text, as well as on the

basis of a sample of spontaneous speech. Furthermore, atomistic scoring is not a very good alternative for holistic rating. It cannot predict a general impression of the quality of a testee's pronunciation. Besides segmental features and stress or intonation patterns there must be one or two other relevant factors in it that determine the quality of pronunciation.

Natural rhythm, liveliness, speaking rate? We do not know yet exactly what they are. Further research has to be carried out.

BIBLIOGRAPHY

Bolus, R.E., F.B. Hinofotis, and K.M. Bailey 1982. An information to generalizability theory in second language research, Language Learning 32, 245-58.

Brennan, R.E. 1983. Elements of generalizability theory. Iowa City: American College Testing Program.

Clark, J.L.D., and S.S. Swinton 1980. The test of spoken English as a measure of communicative ability in English-medium instructional settings. TOEFL Research Report 7. Princeton, N.J.: ETS.

Clark, J.L.D. 1972. Foreign language testing: theory and practice. Philadelphia: CCD.

Gardner, R.C., and W.E. Lambert 1972. Attitudes and motivation in second language learning. Rowley, Mass.: Newbury House.

Harris, D.P. 1969. Testing English as a second language. New York: McGraw-Hill.

Heaton, J.B. 1979. Writing English Language Tests. London: Longman.

Madsen, H.S. 1983. Techniques in testing. Oxford: Oxford University Press.