DOCUMENT RESUME

ED 266 643 FL 015 457

AUTHOR van Weeren, Jan
TITLE Language Testing: Making a Decision on the Basis of Observed Behaviour and "Tapped" Knowledge.
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 *Evaluation Criteria; Foreign Countries; Immigrants; *Language Proficiency; *Language Tests; *Second Language Learning; Teacher Qualifications; *Test Use

ABSTRACT

It seems possible to evaluate language proficiency as behavioral ability by: (1) observing authentic language behavior, (2) examining the tacit knowledge that underlies language behavior or testing specific tasks based on capabilities that are linked to implicit knowledge of a language, and (3) testing the acquired explicit knowledge, such as use of grammatical rules. A test was administered to foreign nationals in the Netherlands who were applying to teach their native languages and cultures to preschoolers. The test, intended to measure the first two factors, consisted of an oral proficiency test of authentic language behavior, a standard multiple-choice test, a multiple-choice test of orthography and morphology, and a cloze test. Analysis of the data found similar results for measurement of language proficiency by evaluating language behavior and by testing implicit knowledge. Problems with the orthography subtest emphasized the need to consider language background when evaluating proficiency. (MSE)

LANGUAGE TESTING: MAKING A DECISION ON THE BASIS OF OBSERVED BEHAVIOUR AND 'TAPPED' KNOWLEDGE

Jan van Weeren (Cito, Netherlands)

0 Introductory remarks

In this paper I will present a rather general framework concerning the linguistic background of language testing. I will stress the fact that a test is not just a set of valid items, but that it is essentially an instrument with a specific function. Various aspects of tests will be discussed, such as the problem of gathering relevant information from testscores, in order to make decisions. A test for immigrant teachers is taken as an example. The results of this test will be related to the general framework presented at the beginning.

1 The nature of language proficiency

By saying that someone is proficient in a foreign language we imply that he or she is able to perform in that language, is able to show actual foreign language behaviour, in short, that he has a certain behavioural ability. There are several ways to look at this ability when it is necessary to measure language proficiency. It can be measured on the basis of actual authentic language behaviour, by determining the quality of specimens of authentic language behaviour. But it is also possible to consider this behavioural ability as a form of tacit, implicit knowledge which underlies authentic language behaviour, but which does not necessarily need to be evaluated through this behaviour. One can specify which tasks someone must be able to perform on the basis if his assumed tacit knowledge of a language, without these tasks necessarily taking the form of authentic language behaviour. In a third approach to this behavioural ability evaluation takes place by virtue of the individual's explicit knowledge about a language. A testee is expected to be able to state the rules concerning the use of the simple present and the progressive form in English, to rattle off the paradigma je suis, tu es, il est, nous sommes, vous êtes, ils sont in French, or to indicate with whom he or she is allowed to be on familiar terms in German. This explicit knowledge was traditionally considered as an important precondition for writing in a foreign language. In short: Evaluation of language proficiency as behavioural ability seems to be possible, firstly by observing authentic language behaviour, secondly by an examination of the tacit knowledge that underlies language behaviour and thirdly by testing the acquired explicit knowledge. In the following I will discuss these approaches successively.

1.1 Evaluation by means of authentic language behaviour

Evaluation of language proficiency as a behavioural ability by means of authentic language behaviour entails a method which is extremely face-valid. If you want to know if a person can take part in a foreign language conversation, you make him or her participate in it. If you want to know if he or she is able to read a text of a specific type, you put questions to him that test his comprehension of the text, or, alternatively, you make her write a précis. As a rule, there will be no objections to such an approach. It is clear and obvious that the procedure makes sense. Add to this that the method entails a positive backwash effect: as a result of the way of testing matters are trained in school that are actually required outside school.

However, when the testing expert comes in a number of problems will arise that increases the more the language behaviour observation resembles the language behaviour that we meet in practice. Human beings usually do not read texts in order to answer a fixed set of reading comprehension questions, but in order to realize specific goals that can vary with the individual and the situation. Their reading comprehension can express itself in various ways. Distinct aspects of the text may be relevant for some, but not for others.

However, testing will require a certain amount of standardization. A conversation which is truly free and open is known for the sort of evaluational problems that are related to the reliability of rating. Regarding the free conversation there is yet another source of unreliability, that is: the testees themselves. Not everyone will be able to speak just as easily about any topic. One should consider factors as empathy and affective thresholds, factors that have very little to do with language proficiency (Underhill 1983).

Another problem is that of content validity: to what extent can observations of incidental language behaviour provide us reliable and complete information about the testee's behavioural ability, that the tester is to judge eventually?

1.2 Evaluation by means of tasks based on implicit knowledge

Specific tasks based on capabilities that are linked up with implicit knowledge of a language can be derived from traditional generative linguistics. This discipline operates on the basis of the concept of competence. Linguistic competence is generally defined as the individual's knowledge of the structure of his language. On the basis of this knowledge a native speaker is able to make judgements about his language system resulting from his 'true linguistic intuitions'. He is able to discriminate between well-formed and not well-formed sentences and to distinguish semantical differences and similarities in sentences that belong to his language.

Dell Hymes expanded the Chomskyan concept of linguistic competence to communicative competence. Competence remains underlying knowledge, but is expanded to include all the aspects of knowledge that affect communicative behaviour. On the basis of this knowledge a native speaker is able to judge not only the

grammaticality of sentences, but also the appropriateness of utterances. His knowledge does not confine itself to the language as a grammatical system, but is expanded to the use of the language. The native speaker knows whether and to what degree something is appropriate, that is, adequate, fortunate and successful in relation to a context in which it is used and evaluated.

Tasks that invoke capabilities linked up with implicit knowledge are very common in traditional language tests, not so much because of the fact that these tests are explicitly based on the theoretical implications of the concept of competence, as because of the necessary concessions one has to make on test-theoretical grounds if language behaviour is to be tested. I will give you two examples of this type of task:

figure 1

HERE'S _____ IMPORTANT INFORMATION FOR ALL OUR CUSTOMERS

1) AN
2) ANY
3) EVERY
4) SOME

One might say that this item is testing the implicit knowledge of a selectional restriction rule (Modelltest VHS Zertifikat English, item 46).

figure 2

1) VOUS ACHETEZ SOUVENT CE JOURNAL ?
2) EST-CE QU'ELLE ACHÈTE LE JOURNAL ?
3) QUAND ACHETEZ-VOUS CE JOURNAL ?
4) VOUS ACHETEZ SOUVENT DES REVUES ?

All utterances are well-formed, but only one is appropriate in this context; the testee must possess some communicative competence (Standardprov. French, Lindblad 1983, p. 57).

From a theoretical point of view the concept of communicative competence is too hybrid, although the concept of linguistic competence cannot be said to be theoretically uncomplicated, either. From a pragmatical point of view, however, the concept of competence gives support to many traditional language tests.

The main question is how test data obtained by means of items based on such tasks, relate to the behavioural ability that we aim at.

It can be hypothesized that the attained level of performance on these tasks gives us an indication of someone's language proficiency as behavioural ability. To put it in a different way: the obtained performance data will represent some measure of language proficiency.

A testing procedure which is based on tasks that a testee must be able to perform in virtue of his competence shares the problem of content validity with the evaluation by means of authentic language behaviour. Every testing procedure has to confine itself to a sample of possible performance. It is necessary to indicate or to test empirically to what extent such a sample is representative.

The hypothesis that testing based on tasks fitting the concept of competence on one hand and the evaluation of authentic language behaviour on the other hand will yield equivalent information about language proficiency as behavioural ability, can be tested empirically. This can be done by determining the concurrent validity of two alternative testforms. Clark (1972) found, for example, correlations from .82-.92 between the FSI-interview and a battery of objective tests for vocabulary and structures.

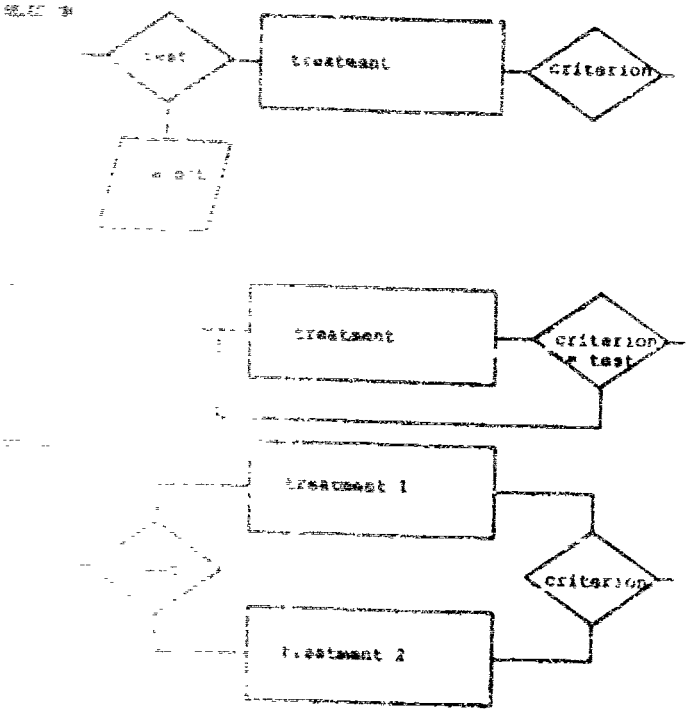
Tests consisting of tasks that the testee must be able to perform because of his implicit knowledge of a language, have the advantage of the possibility of an objective form. An exclusive use of this type of test in instructional settings, however, will inevitably carry with it the disadvantage of an undesirable backwash affect, unless particular measures concerning the curriculum have been provided against this effect. Otherwise educational activities will focus on learning tasks of an abstract nature and training of the actual use of the language will be neglected.

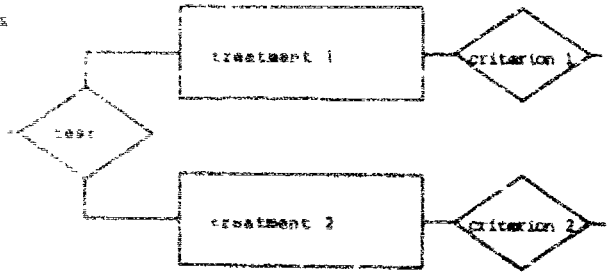
1.3 Evaluation by testing explicit knowledge

In respect to the form of testing that focuses on explicit language knowledge we can be brief. As early as half a century ago it was stated that knowing about a language and knowing of a language are two different things. Explicit knowledge of rules is neither a necessary, nor a sufficient condition for a successful and all-round language use. The first is proven by native speakers of a language who did not receive any formal schooling in linguistics, the second by grammar school pupils who, though having a thorough command of explicit rules, cannot speak a word of Latin or Greek. A teaching process that prepares for this kind of testing should be considered as philology, rather than as language teaching.

2 Test functions

It is remarkable that among applied linguists the form and theoretical background of test items are the main topics under discussion. They mistakenly call a set of test items a test, whereas a test is essentially a different thing, namely a measuring instrument that provides useful information to be used in decision making. In this respect linguists are a bit like instrument makers who talk about a clinical thermometer and think that the thermometer is good if the column of mercury is in different positions with different persons and in the same position with the same person under conditions of repeated measurement. However, a clinical thermometer is only fit if it provides information that enables the user to decide if he or she should stay in bed, should consult a doctor, should take medicines or has reached a period of fertility.
In the same way a test must provide information that enables us to decisions of the following kind:





The concrete test I would like to discuss meant to support a process of making decisions of the first kind. The testees were groups of teachers who had come from various countries: Turkey, Morocco, Spain, Italy, Jugoslavia, Portugal, Greece and Tunesia, to teach their native language and culture to groups of children with the same cultural background in primary schools. The treatment offered was a special course that would make them fully qualified teachers in primary schools. It was observed that these teachers were not integrated into the teaching staffs, mainly because of the fact that they were not qualified to teach any other subject but their native language and culture. Neither were they allowed to teach other pupils than those of their own culture. In order to further the integration and to widen the scope of these teachers the course would give them the oppurtunity to acquire a full qualification. The criterion in the flowchart would consequantly cover the objectives of a regular college of education.

To set an entrance level it was required that the candidate members of the course had a satisfactory oral command of the Dutch language so that they could communicate rather fluently.

On the authority of the inspectorate this requirement was tightened up. To a certain extent command of the Dutch language was required at the level of pupils in the last form of higher level secondary schools. The inspectorate had its reasons for this: the course would only take two years and after finishing it these teachers would be formally qualified to teach the Dutch language to Dutch children in primary schools!

On the basis of these requirements global selectional criteria were set.

Cito, that is the Dutch National Institute of Educational Measurement, was charged with their operationalization.

3 Choice and development of the instrument

With the theoretical considerations put forward at the beginning of this paper in mind it was tried to obtain information about the language proficiency of the testees in two different ways: firstly by eliciting and evaluating authentic language behaviour and secondly by administrating items that would measure the underlying implicit knowledge.

3.1 The oral proficiency subtest

For the evaluation of language behaviour the following type of oral proficiency test was designed. It consists of about 15 stimuli requiring the testees to play a certain part in a situational dialogue. The stimuli are presented verbally. In most of the cases a response is prompted by means of a picture, for example:

AT THE POLICE STATION

One day you discover your wallet is missing. You go to the police station.
Good morning, an officer says. What can I do for you? /.../ Could you tell me what was in your wallet? (picture 1) /.../ Have you any idea where you could have lost your wallet? (picture 2) /.../ etcetera.

These stimuli are followed by circa 15 general questions about their life in Holland.

The responses were each judged on intelligibility/appropriateness and correctness. To each response a maximum of 6 points could be assigned. The maximum meant that a response was perfect: appropriate in the context given, perfectly intelligible and completely correct in a grammatical sense. 5 points were assigned if the response was intelligible and appropriate, although some minor mistakes were made. If it took some effort to understand the response as a result of certain important lexical or grammatical mistakes, 3 points were assigned. If a response was either not forthcoming or not intelligible or not appropriate at all, 1 point was assigned. In all cases one point was deducted if repetition of the stimulus was necessary.
In this rating system a sum total of 180 points with 30 items means that all responses are perfect: intelligible, appropriate and completely correct. There is a rapid decrease in degree of correctness: a score of 150 points means that the average response is appropriate and intelligible, but no more than that. With lower scores a general appropriateness and intelligibility is preserved for a relative long period. Understanding takes some effort with scores below 140. With a score of 120 or 110 points severe misunderstandings will arise and communication will start to break down.

Testing procedures like this one based on authentic language behaviour, have the additional advantage of enabling us to apply natural, that is to say, common sense criteria when cut-off-scores have to be determined. Criteria such as 'oral command of Dutch on a near-native level' or 'being able to make oneself understood' can almost immediately be translated into a score on the oral proficiency test. An example can be taken from the rating system of the FSI-interview: a natural criterion related to overt language behaviour corresponds directly to a specific test score.

In the case of the teacher and the test mentioned above the requirement of oral proficiency was operationalized as a sufficient degree of appropriateness and intelligibility in oral expression as measured by the test. With some safety margin the cut-off-score was set at 145 points.

3.2 Testing implicit knowledge

For the measurement of implicit language knowledge three subtests were composed. One of them aimed at vocabulary and structures: at the implicit knowledge of rules concerning the derivation of words, synonymy, the use of conjunctions and the building of sentences. It consisted of multiple-choice items of a rather straightforward type. The multiple-choice items in the second subtest focused on the orthography of changeable and unchangeable words. This part leaned heavely on the quality of literacy training in Dutch. Therefore it was more like an achievement measure than an indicator of language proficiency. The third subtest consisted of a cloze test in multiple-choice-format. In fragments taken from transcriptions of newscasts every 5th or 7th word was deleted.

Distractors were obtained by administrating an open version of the test with foreign university students. With regard to the underlying knowledge measured by this cloze test it can be concluded that widely different aspects were involved. Apart from lexical fluency the test required from the testees that they could process divergent structurel and logico-semantical information in the context.

3.3 Setting a cut-off-score for tests based on implicit knowledge

There are no natural, that is, common sense criteria for the determination of a cut-off-score in tests that are focused on linguistic and communicative competence. It is not possible to formulate such a criterion in terms of observable and functional language behaviour testees are expected to show. For lack of natural criteria the setting of a cut-off-score will depend on the test itself. In consequence, the outcome is essentielly arbitrary. I will sketch two current standard procedures. According to the first method a group of testees sets its own norm. After the administration of a test the mean score of the group is calculated and the cut-off-score is set somewhere below the mean (for example at a distance of twice the standard error of measurement). In the second method a group of experts is to determine what they consider as a passable or not passable perfomance on the test. This can be done in an informal way, by discussion and consensus, or following to some formal procedure (for example that of Nedelsky, Angoff or De Groot). A combination of both methods is applied in centralized final examinations in Dutch secondary schools. A preliminary cut-off-score is set on a intuitive basis. Depending on concrete testresults an adjustment is made if necessary. The percentages of pupils that pass or fail should approximately equal those of previous years. This is a proper method if testpopulations are sufficiently stable each year with regard to their educational background.

3.4 The choice of a reference population

In some cases there may be an opportunity to set a norm in an empirical way by introducing a well-defined reference population. To give an example, if the language proficiency of a target population should equal the proficiency level of the average

succesful foreign university students, this norm can be obtained by pretesting.
This solution presented itself with the test for the foreign teachers. I remind you of the fact that not only a functional oral command of the second language was required, but also some practical knowledge of Dutch on the level of pupils in the last form of higher level general secondary schools. These are five-form schools following primary education for pupils between the ages of 12 and 17. These provided us with a clear-cut reference population.

The intended norm was defined as the score that anyone could reach who was good enough to reach the last form. At first sight finding this norm seemed very easy. However, we were confronted with a slight complication.

Pretesting would depend on voluntary participation. The test would not have any consequences for the pretest population. Thus there was a change that some pupils would take it less seriously. Apart from sabotage we had to consider the fact that one pupil or the other would have an off-day. For that reason it was rather dangerous to define the norm as the lowest score of the selected pretest population. Therefore it seemed reasonable to take that score as a norm that was two standard deviations below the mean score. With the expected distribution of scores this norm would be reached by circa 95 percent of the pupils.
A pretest was carried out with the following results:

RESULTS OF THE REFERENCE POPULATION (1983)

	n	KR20	$\bar{p}$	$\bar{x} < M-2_{\sigma}$
vocabulary and structures	111	.50	.94	2,7
orthography	111	.86	.89	6,3
cloze	249	.92	.91	4,4

The high reliability coefficients for orthography and cloze gave food for thought. It might very well have been that some pupils did not do their utmost. The item-analysis of the cloze test revealed that several items were skipped at the end, that is, no answers were given. If this occurs systematically, that is, if the same testees skipped each of these items, reliability is flattered. Pretest results of a similar test in 1984 are more realistic from this point of view:

RESULTS OF THE REFERENCE POPULATION (1984)

	n	KR20	$\bar{p}$	$\bar{x} < M-2_{\sigma}$
vocabulary and structures	153	.45	.96	0
orthography	193			
cloze	193	.63	.98	0

Still, reliability of the cloze test is quite high, in view of the impressive percentage of correct answers. However, extremely low scores did not occur anymore.
After the pretesting of the reference population norms were set following the procedure described.

4 Testresults

Administration of the test with 87 immigrant teachers yielded the following results:

	number of items	KR20/α	β
vocabulary and structures	10	.75	.55
orthography	20	.86	.56
cloze	100	.96	.52
speaking prof.	30	.96	.57

Obviously the speaking proficiency test was much easier than the other subtests. No wonder, in view of the fact that only a Basic Interpersonal Communication Skill (Cummins 1979) was involved, without the requirement of native speaker proficiency.
The rating of the elicited responses on the speaking proficiency test was carried out by rotating pairs of raters. Each rater assigned his scores independently. The inter-rater-reliability was surprisingly high. To a certain degree this could be explained by the heterogenity of the testpopulation.

pairs of raters	R
A/B	.98
A/C	.98
B/D	.97
B/C	.96

The following correlation coefficients between the various subtests were found:

cloze x vocab. & struct.	.76
cloze x orthography	.74
cloze x speaking prof.	.71
vocab. & struct. x orthography	.70
speaking prof. x vocab. & struct.	.68
speaking prof. x orthography	.65

5 Conclusions and discussion

On the basis of these modest results we could not reject the hypothesis that the measurement of language proficiency can be done by evaluation of language behaviour as well as by the testing of implicit knowledge. Instruments of both types do provide equivalent results. How these instruments relate to each other may appear from the following survey:
9 candidates with a sufficient score on the speaking proficiency test passed the vocabulary and structures subtest as well as the cloze test. Only one of the candidates scoring just below the cut-off-score of the speaking proficiency test (scores ranging from 140 to 144 points) passed both the vocabulary and structures subte c and the cloze test. Another candidate passed the cloze test only. Below 140 points just one candidate passed both subtests. If we had based our selection decisions exclusively on the subtests vocabulary and structures and the cloze test, only three misclassifications would have resulted in that three candidates out of 87 with an insufficient speaking proficiency would have passed the test.

With regard to the possibility of misclassifications the results of the subtest orthography were more problematic. Of the candidates that passed the speaking proficiency test 16 passed the orthography test as well, but among the candidates just below the cut-off-score there still were four of them that passed the orthography test, and with lower scores there even were six!

It is obvious that orthography is linked up with language knowledge, but to a large extent it can be acquired as an isolated system, especially if it is confined to general rules. Those who passed the orthography subtest with a deficient speaking proficiency, might have acquired their knowledge by their unflagging energy in language classes.

This illustrates the necessity to take the factor language background into consideration when testing language proficiency (Cziko 1984). Language background refers to the type of contact the testees have had with te second language and the opportunity they have had for acquiring the various aspects of the language. If this language background consists of a language course where writing skills are emphasized, language proficiency might be flattered if these skills are much represented in the evaluation procedure. If, on the other hand, the various aspects of a language are trained in a more balanced way, distinct subtests will represent language proficiency as a whole more adequately.

Bibliography

Hymes, D. (1972), 'On Communicative Competence', in J.B.Pride & J Holmes (eds.), Sociolinguistics, Harmondsworth, 269-93.

Underhill, N. (1982), 'Commonsense in Oral Testing: Reliability, Validity and Affective Factors', in E.W.Stevick a.o. (eds.), On TESOL '82: Pacific Perspectives on Language Learning and Teaching II, Tesol Washington, 125-39.

Backlund, P. & J.Wiemann (1978), 'Current Theory and Research in Speech Communicative Competencies: Issues and Methods', Paper presented at the annual convention of the American Educational Research Association, Toronto.

Cziko, G. (1984), 'Some Problems with Empirically-based Models of Communicative Competence', Applied Linguistics, Vol. 5, no.1, 23 38.

Lowe, P. (1976), The Oral Language Proficiency Test, Interagency Language Roundtable, Washington D.C.

Deutscher Volkshochschul-Verband, Pädagogische Arbeitsstelle (1980), Das VHS-Zertifikat Englisch, Bonn.

Lindblad, T. (1983), 'The Swedish System of Final Examinations in Modern Foreign Languages', in J.van Weeren (ed), Practice and Problems in Language Testing 5, Cito Arnhem, 51-63.

Cummins, J. (1979), 'Linguistic Interdependence and the Educational Development of Bilingual Children', Review of Educational Research 49, 222-51.