DOCUMENT RESUME

ED 282 413 FL 016 704

| | |
|---|---|
| AUTHOR | de Jong, John H. A. L. |
| TITLE | Testing Foreign Language Listening Comprehension. |
| PUB DATE | 82 |
| NOTE | 11p.; Paper includes a table and six figures that contain small print. |
| PUB TYPE | Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150) |
| EDRS PRICE | MF01/PC01 Plus Postage. |
| DESCRIPTORS | Cloze Procedure; English (Second Language); Evaluation Criteria; Foreign Countries; Item Sampling; Language Proficiency; *Language Tests; *Listening Comprehension; Native Speakers; Objective Tests; Pilot Projects; Postsecondary Education; Second Language Learning; *Second Languages; *Test Construction; Test Format; Test Items; Test Reliability; *Test Validity |
| IDENTIFIERS | *Netherlands |

ABSTRACT

The development and validation of a test of listening comprehension for English as a second language at the Dutch National Institute for Educational Measurement (Cito) is described. The test uses two distinct item formats: true-false items and modified cloze items with two options. Both item formats were found to measure foreign language listening comprehension in a valid and reliable way, testing comprehension of a large variety of language samples. However, the cloze item appeared to demonstrate better psychometric qualities. Further testing of the true-false item format is recommended. (MSE)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED282413

# TESTING FOREIGN LANGUAGE LISTENING COMPREHENSION

John H. A. L. de Jong

National Institute for Educational Measurement, Cito, Arnhem, The Netherlands

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

de Jong

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

FL 016 704

BEST COPY AVAILABLE

**Testing Foreign Language Listening Comprehension.**

John H.A.L. de Jong, National Institute for Educational Measurement, Cito, Arnhem, The Netherlands

## Introduction

The present paper reports on a pilot test of listening comprehension of English as a foreign language at the pre-university level. The test has recently been developed at the Dutch National Institute for Educational Measurement (Cito) in a research project designed to study new methods of testing listening comprehension.
I will first briefly go into the question, why it was decided to develop the test, mention the objectives of the test, and then give a description of the test itself.
In the body of the paper I will deal with a number of aspects of the validation procedure for the test. I will not deal with aspects such as concurrent validity or KR20 reliability. First of all because they demand the presentation of long tables of data, a procedure hardly efficient at a conference. Data on these criteria have been documented, however, in the official report on the project (Jong, de, en van den Nieuwenhof 1982). Secondly, because some aspects I do wish to deal with here offer the possibility of presenting new ways of looking for validity, which is quite close to one of the topics of this symposium.

## Why a new listening comprehension test?

For almost ten years, tests of foreign language listening comprehension have been constructed at Cito. The tests were originally developed in a research project at the University of Utrecht in the Netherlands (Groot, 1975). Annually a new test is constructed for English, French and German at four proficiency levels defined by the four different types of secondary education in the Netherlands.
A validation research on these tests undertaken at Cito in the late nineteen-seventies reported severe constraints imposed by the multiple choice technique used exclusively in the tests (van den Nieuwenhof et al. 1979). One of the constraints was the choice of language samples: only samples permitting a one-phrase summary of the global contents could be used.
A second constraint was on the difficulty of the language samples. The multiple choice technique offers so much help to the testees, that in order to prevent the scores to go up to high, and thereby reducing variance in test scores, language samples had to be selected at a difficulty level too high for the testees' proficiency: testees would not be able to understand the samples without the support offered by the multiple choice questions. Furthermore, no relation could be found between the actual difficulty of the language samples as observed in open ended questions, and the difficulty indices of the corresponding multiple choice items. It was therefore concluded, that other item formats needed to be developed in order to ensure the validity of Dutch national listening comprehension tests.

## Objectives

The new item formats would have to be used in a pilot test in order to permit their evaluation. The pilot test would have to make use of samples from various types of language use and of item formats different from the traditional multiple choice items. The test would have to permit objective, machine scoring if only because of the number of testees -over 350.000- that takes foreign language listening comprehension tests each year.
The test would have to be acceptable to its users: teachers and students.
The test would have to yield a reliable and valid measurement of foreign

language listening comprehension.
For practical reasons one language at one proficiency level was chosen for the pilot test: English at the pre-university level.

The test itself

Language samples in the pilot test are from three different sources:
- recordings in the series 'Topical Tapes', a special B.B.C. overseas service, the copyrights for which Cito obtained for the Netherlands;
- American and English news broadcasts recorded from the radio;
- interviews with native speakers recorded by Cito.

Two item formats are used:
- true-false items
  Testees are to decide whether the statement printed in the test booklet is in accordance with what is said on the tape.
- modified cloze items with two options
  Words to be deleted from the text are not randomly selected, but are chosen for their semantic relevance in the context. In each sample on the tape one word -or group of words- is cut out from the tape and replaced by an electronical sound. Testees are to decide which of the two options presented in their test booklet can be used to restore the text. Thus a typical problem with cloze items -acceptable word or exact word scoring- is avoided and at the same time one avoids testing a productive skill.

Four different types of language use are included in the test: discussions, telephone conversations, news programmes and interviews. These are combined with the two item formats in such a way as to form six different subtests. The subtests are divided over two tapes to enable testees to take the test in two separate sessions. After editing the preliminary version of the test, 132 items were selected for the final version.
Each item consists of a language sample of about 20 seconds on tape and a printed question, which testees are to answer in a ten second pause provided for on the tape. Testees listen to each sample only once.
Figures 1 and 2 present an example of both item formats in the test.

Figure 1: True-false item from the pilot test

Item no 17

Tape:

(Wilson:) One of the things that worries me about them is how people are going to be able to keep track of their spending when they've just got a plastic card which they hand to the shop keeper. Now...

(Portescue:) You've just raised a very important question, that if you're paying for goods with that wad of notes you can see how many you've spent and you have some idea of the total of your day's shopping whereas if you go into a number of eh shops and pay with eh the same credit card, you come away with nothing other than receipts and there's no sense of actually spending money.

Question booklet:

17 The use of a credit card can make people less aware of the amount of money they spend.
A True B False

Figure 2: 'Cloze' item from the pilot test

Item no 43

Tape:

Prince Charles has been speaking about Britain's industrial future. He told delegates at a conference in Bournemouth of the iron and steel trades union that if only the two sides of industry would co-operate more, Britain's performance internationally could be ---Buzz---

Question booklet:

43 A vastly improved
B very disappointing

## Evaluation

Information on the test was gathered by means of
- a questionnaire on the test objectives presented to foreign language teachers;
- try-out sessions of the test on two independent samples of the target population and on a group of native speakers of comparable age and educational background.

### Face validity

Although face validity is not a psychometrical value it does play an important role in testing. Neither teachers nor students will take the administration of a test seriously if they fail to see a relation between the test and what it is supposed to measure.
A questionnaire on test objectives was sent to 150 Dutch foreign language teachers who did not know the test and to 25 teachers who had used the test in class. The results are presented in table 1.

Table 1: Objectives for a listening comprehension test according to teachers

| | not acquainted with pilot test n = 145 | acquainted with pilot test n = 23 |
|---|---|---|
| Language use: | | |
| interview | 99% | 100% |
| discussion | 48% | 87% |
| dialogue | 83% | 91% |
| Acoustical circumstance: | | |
| natural surroundings | 32% | 48% |
| telephone | 31% | 65% |
| radio | 58% | 78% |
| studio | 88% | 78% |
| Item formats: | | |
| true-false | 19% | 74% |
| cloze | 40% | 57% |
| multiple choice | 92% | 100% |

The group of teachers that did not know the pilot test showed a preference for test characteristics that most resembled those of the traditional listening comprehension tests in the Netherlands: samples of interviews recorded in sound studios accompanied by multiple choice questions. A large

minority of this group would, however, appreciate a more general sampling of language use (e.g. discussions, dialogues, documentaries) in less favourable acoustical circumstances (e.g. radio, telephone, natural surroundings) accompanied by a larger variety of item formats (e.g. true-false items, cloze items). The findings for the group that had used the pilot test were quite different. The majority of this group would expect a listening comprehension test to present the traditional characteristics to the same extent as the new characteristics they had come across in the pilot test. It was concluded that acquaintance with the test was a relevant factor and that more variation in language use and item formats used in listening comprehension tests would receive the general support of foreign language teachers.

Content validity

One of the problems in assessing language skills is the fact that language represents an infinite universe. It is impossible to define the skill to be measured in an enumerative way. One cannot ensure the content validity of e.g. a listening comprehension test simply by sampling from every possible instance of language use in a listening situation. The Dutch government plainly requires 'an examination of the listening ability in the foreign language' without any further definition of the performance domain. Groot (1975) originally formulated the objectives for Dutch national foreign language listening comprehension tests as the ability to 'understand the foreign language, spontaneously spoken by educated native speakers' and restricted the domain by demanding 'normal conversational tempo' and excluding 'extremely informal elements', 'lexical and syntactic elements that are incomprehensible to less educated native speakers' and 'topics requiring specialistic knowledge'. The performance domain remains rather large and the restrictions are open to different interpretations.
One can, however, evaluate content validity in a relative way: the more instances of real life listening situations can be included in a test the better content validity can be achieved. Each item format brings along constraints on the choice of language samples. Using more item formats is a method to minimize the effect of these constraints and thus to improve content validity. Two item formats new in Dutch national listening comprehension tests were used in the pilot test and as a matter of fact permitted the use of language samples in the test that were inconceivable in the traditional multiple choice tests.

Construct validity

For construct validation it is necessary to define the concept which accounts for performance on the test. In a foreign language listening comprehension test differences among testees in their ability to understand the spoken language in question should account for differences in their raw scores on the test. Without going into the discussion on the Unitary Competence Hypothesis versus the belief that linguistic competence can be broken down into a number of totally distinct factors or Vollmer's suggestion of a hierarchical model (Vollmer and Sang 1981; Vollmer 1981) one can safely assume that native speakers of a language do possess the ability to understand their native tongue. Native speakers, however, differ in their ability to follow the spoken language according to their ability to cope with the language material at the conceptual level. To rule out non-linguistic factors as much as possible the concept that accounts for performance on a foreign language listening comprehension test should be defined as: "The ability to understand the foreign language at the level of native speakers of comparable age and educational background." A group of native speakers thus defined will have to do extremely well on the test and in any case will have to do no less than the most able listeners in the

target population of non-natives. Furthermore no significant variance in native speaker scores on the test is to be expected.
Figure 3 represents the mean score and standard deviation for each of the subtests observed in try-out sessions of the test on a sample (n = 387) from the target population and on a group of native speakers of comparable age and educational background (n = 30). Subtest 2B made use of both item formats, therefore the results are presented seperately as 2B1 and 2B2.

Figure 3: Subtest results of target population and of native speakers

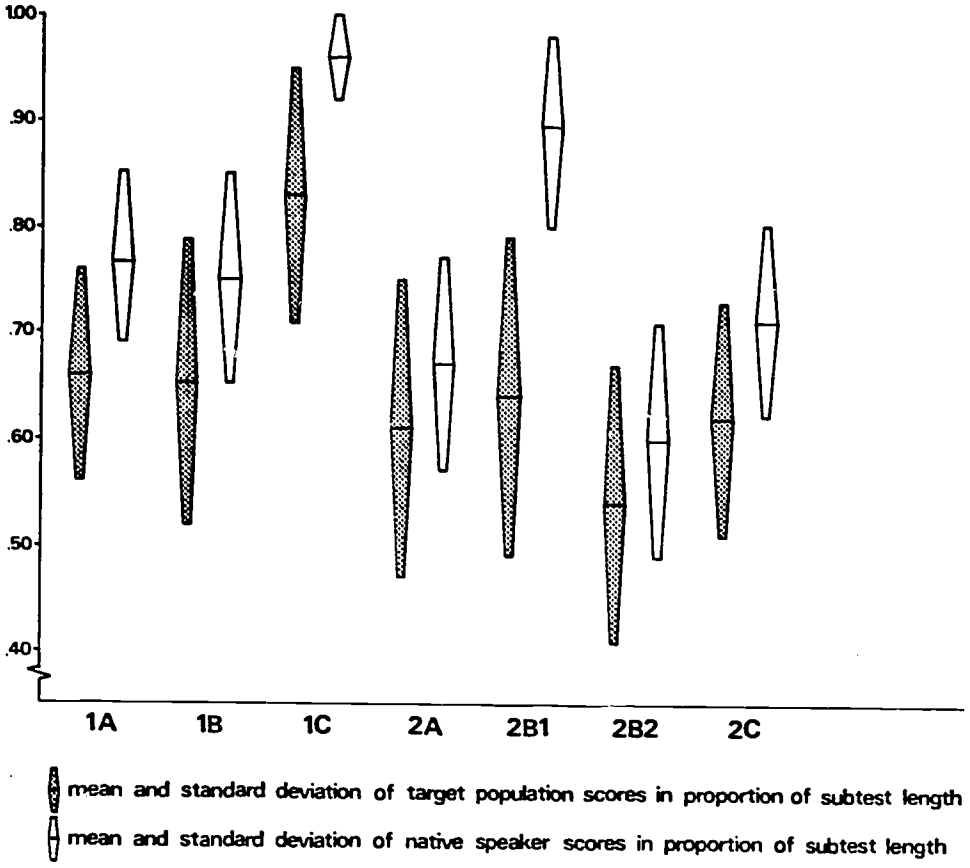
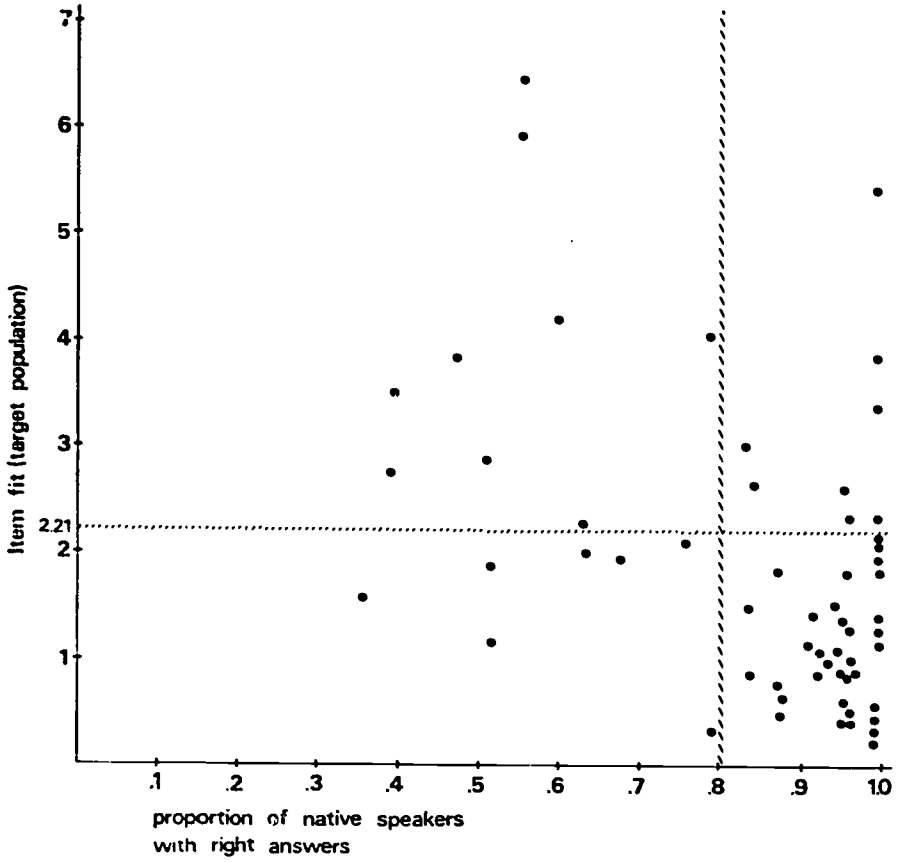



Figure 3 illustrates that native speakers scored higher than the target population on each of the subtests and that standard deviation in native speaker scores is smaller than in target population scores. Differences were checked for significance by means of the Mann Whitney test (Siegel 1956, pp 116-127). For a two tailed test at the 1% level, difference was not significant for subtest 2B2 ($p = .47$). For subtest 2A significance was at $p < .01$ and for the other subtests at $p < .001$. From the figure it is obvious that subtests 1C and 2B1 come up best to the expectations. Both these subtests used the modified cloze items with two options. It seems likely that the extremely poor quality of subtest 2B2 results from the combination of two item formats in test 2B. Teachers reported that testees concentrating on the cloze type items forgot to listen to the samples at the level necessary to answer the more global true-false items.
Rasch analysis was used for the interpretation of test results at the item level. Rasch analyses were done by computer with the program CALFIT (Wright and Mead, 1975). The unconditional maximum likelihood procedure (UCON) of this program was used to estimate ability and difficulty parameters (Wright and Stone, 1979). The Rasch model for test analysis is a so called latent trait model and specifies a relationship between observable test performance and the unobservable traits or abilities assumed to underlie performance on

the test. Furthermore the model postulates unidimensionality of the underlying trait. Because native speakers can be expected to show greater ability in listening comprehension, they should have higher probability of getting the right answer on each item, if the item requires listening ability. Thus correspondence between native speaker ability and the latent trait in Rasch analysis of foreign language learner response would be an indication of construct validity. Figure 4 pictures this relation for part 1 of the pilot test.

Figure 4: Native speaker response (x-axis) versus measure of fit in Rasch analysis of target population response (y-axis)

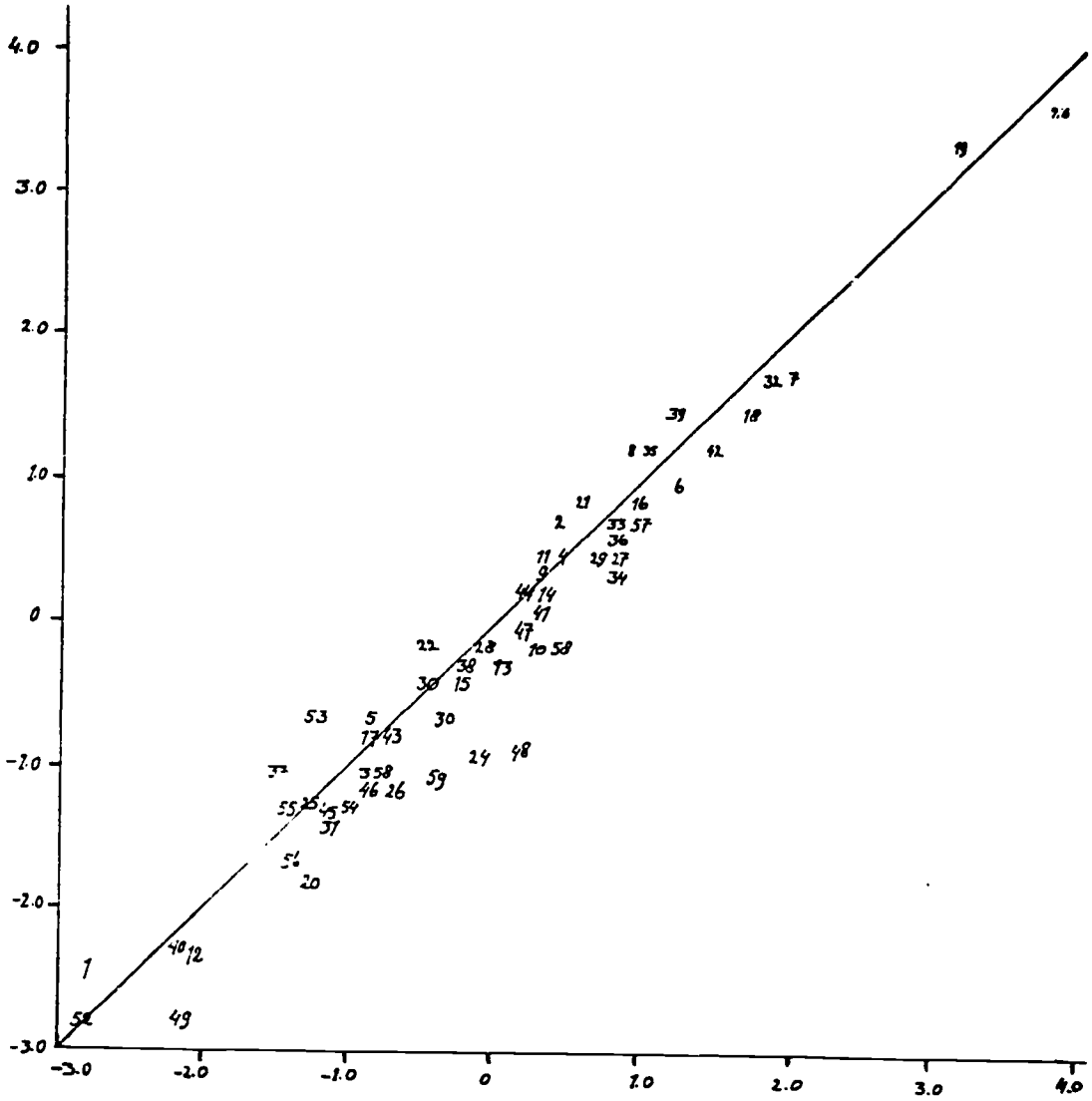



Every dot in figure 4 represents one of the 59 items in part 1 of the pilot test. The critical upper limit of the fit measure at 2.21 (.05 level of significance) is indicated by a horizontal line. A vertical line at .80 on the x-axis represents what is taken as the lower limit for items presenting no difficulties for native speakers. Most of the items (35 out of 59) fall within both limits. Another 9 items surpass both limits and thus also confirm the relation: misfitting items are items that are too difficult for native speakers. For a total of 44 out of 59 items there is a correspondence between native speaker response and the latent trait observed in Rasch analysis of the target population. If the true-false items are analyzed separately this correspondence has been shown to exist in both parts of the pilot test for 54% of the true-false items and 84% of the cloze items (Jong, de, en van den Nieuwenhof, 1982). Evidence for the possibility of improvement of this correspondence by means of an item selection procedure based on the data from Rasch analyses has been reported (Jong, de, 1982).

**Transferability**

A try-out session of part 1 of the pilot test on a second, independent sample from the target population was organized a year after the first session in order to evaluate transferability of the measurement. Only minor differences in test data were observed. The most convincing evidence for the consistency of the measurement is obtained by a comparison of the item difficulty parameters in Rasch analyses of the test results of the two sessions as is presented in figure 5.

Figure 5: Difficulty parameters for the 59 items in part 1 of the pilot test computed from the 1st try-out session (y-axis) and the 2nd try-out session (x-axis) on the target population

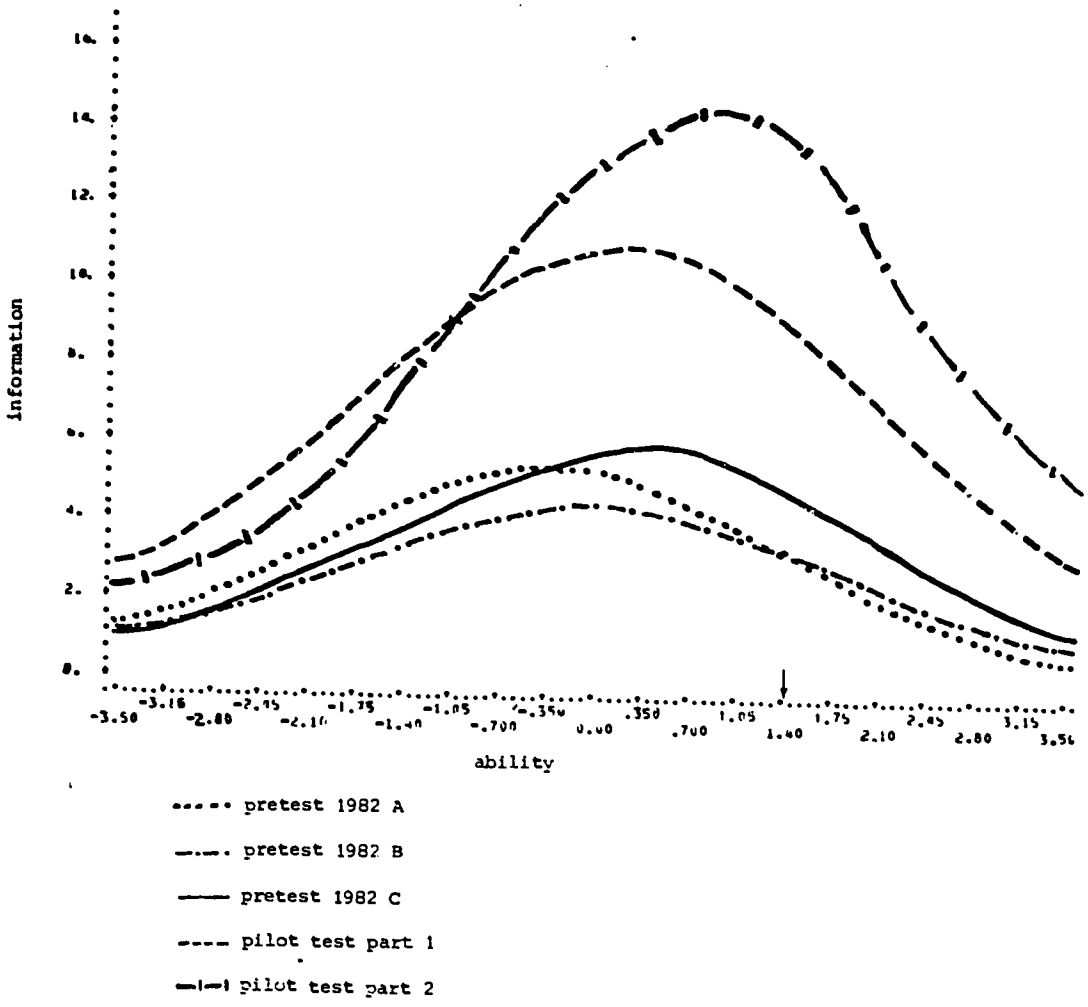



Efficiency

Administration time of the pilot test is at an average of two items per minute versus one item per minute in the traditional multiple choice tests. The number of words testees have to read in their test-booklet is substantially reduced as well: an average of 15 words per item for the true-false items and 3 words per item for the cloze-type items as opposed to 40 words per item for the traditional multiple choice format. Thus test performance is less dependent on reading ability in the pilot test. Figure 6 represents the test information curves of five different tests of equal length in administration time and clearly demonstrates the aspect of

efficiency: both part 1 and part 2 of the pilot test contain more items and therefore supply more information at all ability levels. The mean ability of the target population is indicated in the figure by an arrow.

Figure 6: Test information curves of traditional pretests A, B and C 1982, and pilot test parts 1 and 2




## Conclusion

The pilot test discussed here makes use of two distinct item formats: true-false items and modified cloze items with two options. Both item formats constitute a means to measure foreign language listening comprehension in a valid and reliable way. The item formats allow testing the comprehension of samples taken from a large variety of instances of language use. The cloze type item, however, seems to demonstrate better psychometric qualities. Further investigations of true-false items will be necessary to establish whether their inferior quality compared to the cloze type items must be ascribed to intrinsic characteristics of the item format.

## References


Groot, P.J.M., 1975: Testing communicative competence in listening comprehension; in: R.L. Jones and B. Spolsky, (eds.), Testing language proficiency, Center for Applied Linguistics, Arlington, Virginia

Jong, J.H.A.L. de, 1982: Focusing in on a latent trait: An attempt at construct validation by means of the Rasch model, paper presented at the Fifth International Language Testing Symposium, IUS/Cito, to appear in the proceedings: J. van Weeren, (ed.), Practice and Problems in language testing 5, Cito, Arnhem

Jong, J.H.A.L. de, en H.W.M. van den Nieuwenhof, 1982: Een experimentele luistervaardigheidstoets, Specialistisch Bulletin no 14, Cito, Arnhem

Nieuwenhof, H.W.M. van den, J.H.A.L. de Jong, M.G.J. Boeijen, et al., 1979: Validatie van luistertoetsen moderne vreemde talen, memo 361, Cito, Arnhem

Siegel, S., 1956: Non-parametric statistics for the behavioral sciences, Mc.Graw-Hill, New York

Vollmer, H.J., 1981: Receptive versus productive competence?: Models, findings and psycholinguistic considerations in L2-testing, Paper presented at the Sixth International Congress of Applied Linguistics, AILA 81, Lund (personal communication)

Vollmer, H.J., and F. Sang, 1981: Competing hypotheses about second language ability: A plea for caution, (personal communication), to appear in: J.W. Oller, (ed.), Issues in language testing research

Wright, B.D., and R.J. Mead, 1975: Calfit, Research memorandum number 18, Department of Education, University of Chicago, Chicago

Wright, B.D., and M.H. Stone, 1979: Best test design: Rasch measurement, Mesa Press, Chicago