DOCUMENT RESUME

ED 282 417 FL 016 708

AUTHOR de Jong, John H. A. L.
TITLE Focusing in on a Latent Trait: An Attempt at Construct Validation by Means of the Rasch Model.
PUB DATE May 83
NOTE 26p.; In: van Weeren, J., Ed. Practice and Problems in Language Testing 5; see FL 016 706. Small print in figures may affect legibility.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Ability Grouping; Cloze Procedure; Correlation; *English (Second Language); Evaluation Criteria; Foreign Countries; Language Proficiency; *Language Tests; *Listening Comprehension; Objective Tests; Performance Factors; Second Languages; *Statistical Analysis; Testing Problems; *Test Items; *Test Validity
IDENTIFIERS Netherlands; *Rasch Model

ABSTRACT

The Rasch model for test analysis is a latent-trait model, which specifies the relationship between observable test performance and the unobservable traits or abilities assumed under test performance. In most cases, the test constructor has no clue as to whether the latent traits postulated by the model are indeed the abilities he wants to measure. A study using an English-language listening comprehension test for both native speakers and Dutch students of English as a second language has yielded data that may help to identify the ability that, according to Rasch test analysis, underlies the main trait measured by the test. (Author/MSE)

Focusing in on a Latent Trait: An Attempt at Construct Validation by means of the Rasch Model,

John H.A.L. de Jong

The Rasch model for test analysis is a so-called latent trait model. In a model of this kind, a relationship is specified between observable test performance and the unobservable traits or abilities assumed to underlie performance on the test. In most cases the test constructor has no clue as to whether the latent traits postulated by the model are indeed the abilities he wants to measure. Try-out sessions of a listening comprehension test on native speakers and on foreign language learners have yielded data possibly offering a means to identify the ability which, according to Rasch test analysis, underlies the main trait measured by the test.

Introduction

As the subtitle of my paper indicates I will try to prove construct validity of a test by means of the Rasch model. The test I will deal with is a foreign language listening comprehension test, which has recently been developed at our Institute for Educational Measurement (Cito). I will start with a brief description of the test. Secondly I will define the ability, which we propose to measure with the test, which is in fact the theoretical construct. In the body of the paper I will offer evidence for a relationship between the construct as defined and the measure of fit of items according to the Rasch model. Finally I will illustrate the assumptions made in this paper with examples from the actual test.
After Phiel Theunissen's paper (this volume) I feel free to assume a certain knowledge of the Rasch model but I will explain concepts of particular importance for the evidence which I hope to give.

The test itself

It is important to make the preliminary remark that the test was developed as a research project: to investigate techniques of testing foreign language listening comprehension, techniques which were new to us. Obviously such a test will contain a larger number of items of poor quality than any of our tests constructed for actual use in final examinations.

The test uses life recordings taken from different radio programs, cut into rather short samples of spoken language (about 25 seconds each). The language to be tested is English. Testees listen to the tape, hear each sample only once and have to respond to a multiple choice question with two options printed in a test booklet. They have a 10 second pause in between samples to decide on their choice.

Two types of items were used:

A True-false items
Testees have to decide whether the statement in the test booklet is in accordance with what has been said on the tape.

B Modified cloze items with two options
Words to be deleted from the text were not randomly selected but were chosen for their semantic relevance. In each sample a word or group of words was cut out from the tape and replaced by an electronical sound. Testees were to decide which of the two options presented in their test booklet could be used to restore the text. Thus the typical problem with cloze items - acceptable word or exact word scoring - could be avoided and no productive skill was tested.

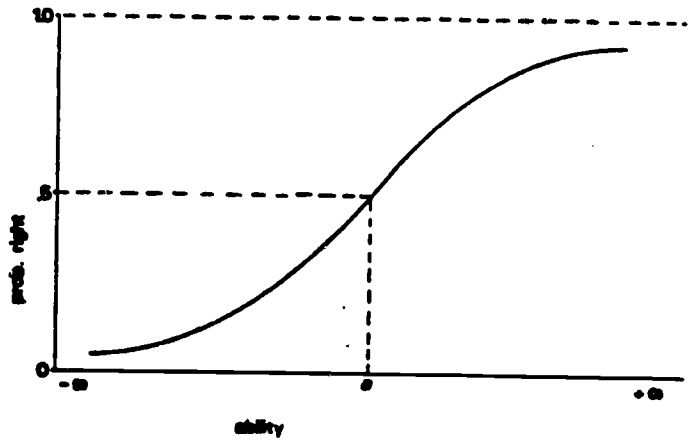
The test was tried out on two groups:
- a representative sample of 575 Dutch pupils taken from the target population (pupils preparing for final Secondary School examinations after 6 years of English as a foreign language)
- a group of 30 native speakers of comparable age and educational background

The construct

For construct validation it is necessary to determine the concept which accounts for performance on the test (American Psychological Association, 1974; Cronbach and Meehl, 1955). In a foreign language listening comprehension test we hope that difference in raw scores on the test can be explained by difference in ability to understand the spoken language in question.
What is the ability to understand a spoken language? 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 this ability.
Still we all know that native speakers differ in their ability to follow the spoken language according to their ability to cope with the language material at the conceptual level.
Obviously, defining foreign language listening comprehension as 'the ability native speakers demonstrate in understanding spoken samples from their native tongue' is too comprehensive. To rule out non-linguistic factors the concept should be narrowed down to: '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 items in the test and in any case, they 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. I will not go into this point here but small variance in native speaker scores on the test has been reported (de Jong en van den Nieuwenhof, 1982).

In terms of the Rasch model: the model postulates unidimensionality of the ability or underlying trait. Because we expect native speakers 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.

Figure 1: Item characteristic curve for item i



Now please consider figure 1 which represents the item characteristic curve for an imaginary item i. Different levels of ability (x-axis) are plotted against the chance of getting item i right (y-axis). A person of ability 'a' e.g. has exactly .50 chance to get the item right. A chance of 1.0 is of course the highest possible chance and would require ability at + ∞. In other words a chance of 1.0 is the upper limit, the ceiling for any person doing item i.

Figure 2: Item characteristic curve for item i

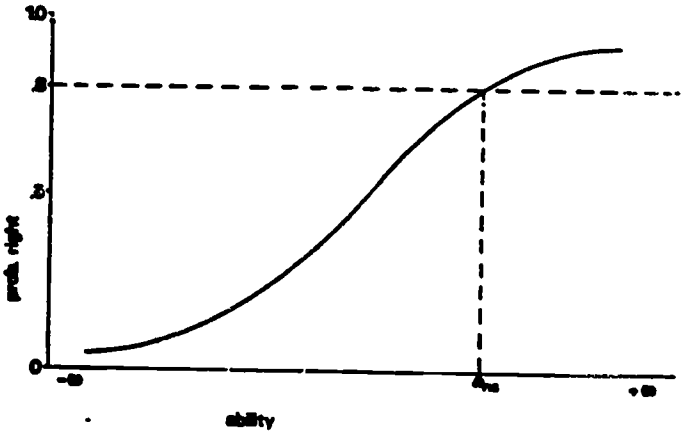
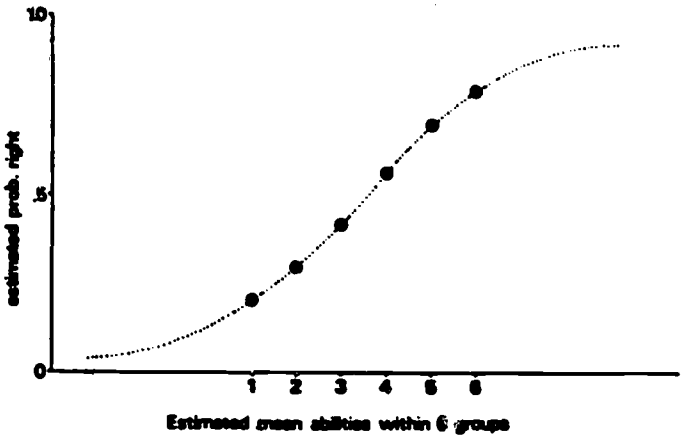


Figure 2 represents this relation between ability and probability of getting item i right again. On the y-axis a dotted line at .8 represents the observed proportion of right answers in a group of native speakers defined as before. From their observed responses we can estimate the

mean ability of this group of native speakers: a_{ns}.
And because we expect non-natives to show less ability in solving the item than native speakers - their ability level on the x-axis should be to the left of that of native speakers (a_{ns}) - we expect them to have less chance in getting the item right. This implies that for the target group of non-native speakers the upper limit of their chance to get the item right is set by the native speaker scores on that item: for this item .8 (and not 1.0). If they score higher than the native speakers then the ability required for the item is not the ability we set out to measure.

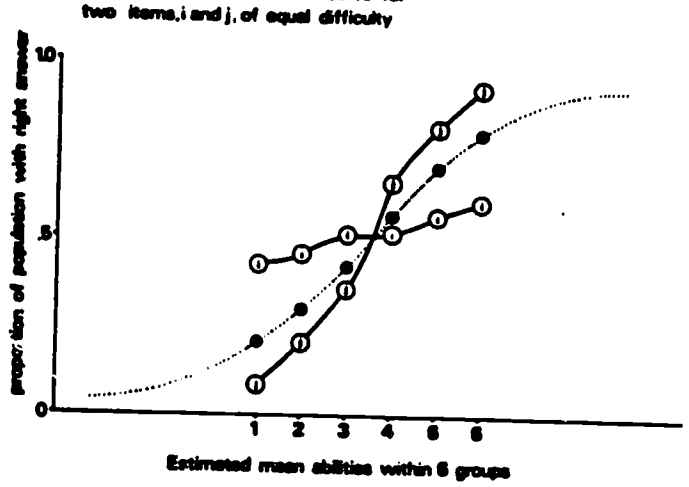
For this research Rasch analysis was done by computer with the program CALFIT (Wright and Mead, 1975). The unconditional maximum likelihood procedure (UCON) of this program was used to estimate ability and difficulty parameters (Wright and Stone, 1979).
The program divides testees into six groups of roughly the same size according to level of ability estimated from their performance on the test.
The program then calculates, for every item, the probability of success on that item for the six groups, and expresses the probability as an expected number of right answers in each group.

Figure 3: Item characteristic curve for item i



In fact this presents us with a rough estimation of part of the item characteristic curve (figure 3). Only a 'rough estimation', because no more than six points (i.e.: at the six ability levels) are defined, and only a 'part' because ability is not likely to vary from minus infinity to plus infinity among the testees.
This estimation of the probability of success is compared to the actually observed mean performance of each of the six groups on the item.
The difference between the estimated and the observed performance is the basis for the calculation of a measure of fit of the item in the test.
Of course there are numerous possibilities for departure of the observed performance from the estimated curve.

Figure 4: Departures from estimated curve for two items, i and j, of equal difficulty



●——● Estimated probability of right answers for items i and j at the ability level of 6 groups.
①—① Observed proportion of groups 1 to 6 with right answers on item i; weak discrimination between groups.
①—① Observed proportion of groups 1 to 6 with right answers on item j; strong discrimination between groups.

In figure 4 two of these possibilities are presented, for two items (i and j) of equal difficulty: the probability of success for each of the six groups is the same on both items. The items do not differ much in degree of departure from the expected curve thus their measure of fit in the test will not differ much. But on item i a larger number of persons from the low ability groups gave a right answer than was expected, whereas in the higher ability groups there are fewer right answers than expected. The item discriminates less well between the ability groups than we would expect from their performance on the whole test. For item j it is just the opposite: this item is highly discriminative.
In fact, the difference in ability between the six groups is invariable, or at least is not likely to change during the administration of the test, thus the differences in performance between the groups should remain constant on the items in the test if the items all measure the same ability.
Now in postulating unidimensionality of the ability, what we expect in general is that whenever an item does not fit, i.e. does not measure the same differences in ability between the groups of the target population as the whole test does, the item requires a different ability and natives will not necessarily show high scores on this item; they might even obtain lower scores than non-natives.

However, imagine a test constructor designing a test to measure physical strength. As items he uses different sizes of nails. He postulates: large long nails require more strength to hammer down than small, thin ones do. He scores the persons taking his test by the number of blows they need to drive in each nail completely. Obviously, if he presents his test to people who are poor at aiming, he

is liable to find physical strength to be higher correlated with number of smashed fingers than with number of blows. What he is doing is measuring two distinct traits or abilities at the same time: physical strength and skill at aiming.

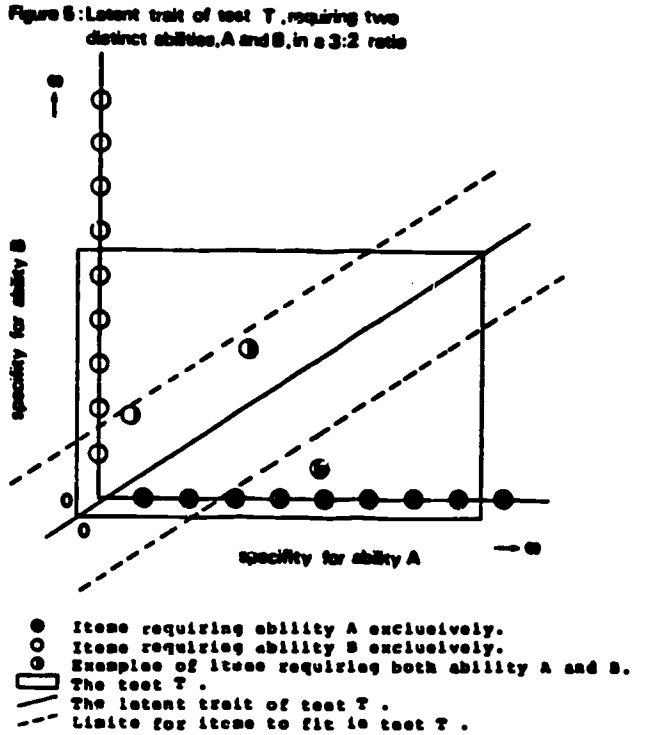


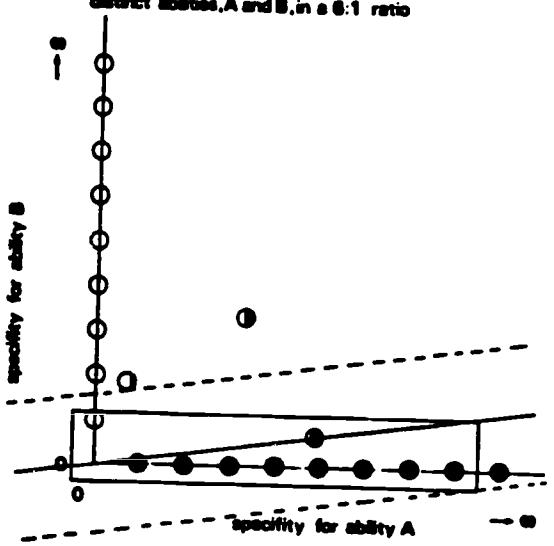
Figure 5 gives a graphic representation of such a test. Test T requires two distinct abilities A and B in a 3:2 ratio. The items requiring either ability A or ability B exclusively are represented as dots on two lines intersecting at an angle of 90°. Items at the intersection of the two lines require no specific ability, neither for ability A nor for ability B. Moving away from the intersection, items require more and more specificly either ability A or ability B and will consequently discriminate more on that ability between persons taking the test. Because the test requires both abilities A and B, the latent trait of test T will be defined as a line going through the intersection of trait A and trait B at an angle defined by the ratio of the two abilities in the test. For items to fit in the test they will have to test both abilities more or less in this ratio or one of either ability at a not too specific level. Items discriminating highly on either ability will not fit.

Because test T contains more items requiring ability A than items requiring ability B, highly discriminating items on the trait of test T are more likely to test ability A specificly, whereas items with low discriminative power in the test have a greater chance to be items testing ability B.

In the example of the test constructor who wanted to test physical strength the best solution for him is to get rid

of the tiny, thin nails that require a good aim. What will happen with the test is shown in figure 6.

Figure 6: Latent trait of test T', requiring two distinct abilities, A and B, in a 6:1 ratio



Items requiring ability A exclusively.
Items requiring ability B exclusively.
Examples of items requiring both ability A and B.
The test T'.
The latent trait of test T'.
Limits for items to fit in test T'.

Test T' still requires the same two distinct abilities, but now the ratio is 6:1. A number of items requiring ability B have been deleted from test T and a new definition of the latent trait for test T' is obtained. The angle formed by the latent trait of test T' and ability A has become smaller and automatically more items requiring ability A will fit in test T'. Items with high discriminative power on ability A now fall within the limits for items to fit in test T'.

With this picture in mind we can redefine our expection mentioned earlier: whenever an item does not fit in our listening comprehension test and this item has a low discrimination index, we do not expect native speakers to do well on this item. In fact, they might just as well score lower than (some of) the non natives from the target population. If, however, an item does not fit, and misfit is due to high discriminative power on the trait of the test, we expect that this item is no problem at all for natives; the item will eventually fit after the items showing misfit in combination with a low discrimination index will have been deleted from the test.
In other words: we expect to find a correspondence between native speaker ability and the latent trait observed in Rasch analysis of foreign language learner response, and to be able to improve this correspondence by deleting items on which low native speaker scores were observed.

Results

In figure 7 the discrimination index of the misfitting items of the listening comprehension test is plotted against the measure of fit of the items. The dotted horizontal line represents the critical upper limit of the fit-measure (.05 level of significance in the F-distribution, with 5 over infinity degrees of freedom). The vertical line represents the median discrimination index: items to the right have more discriminative power than the test in general and items to the left are less discriminative.

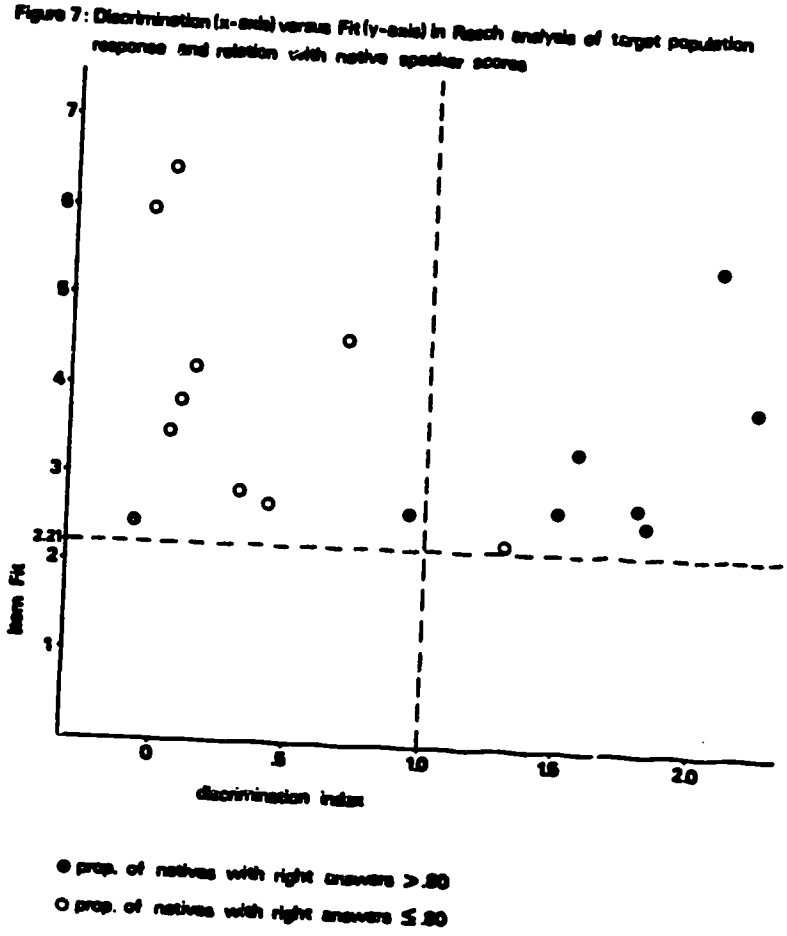


Figure 7: Discrimination (x-axis) versus Fit (y-axis) in Rasch analysis of target population response and relation with native speaker scores

Each dot in the figure represents one of the misfitting items. In addition, for each item an indication is given whether more or less than .8 of the group of native speakers chose the right answer. We take .8 of the native speaker group as the lower limit for items that present no difficulties to the defined group of native speakers. The items tend to form a V-shape: the more the items deviate from a discrimination index of 1.0 the less well they fit in the test. From the picture it is clear that in general highly discriminative misfitting items gave less trouble to native speakers than low discriminating misfits.

Now we feel confident to start the operation of pulling out nails, i.e.: deleting items from the listening comprehension test.

Figure 8 A and 8B: Native speaker response versus measure of fit in Rasch analysis of target population response (59 items)

Figure 8A

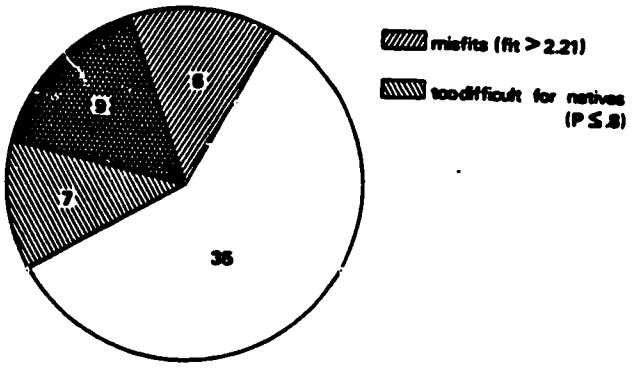
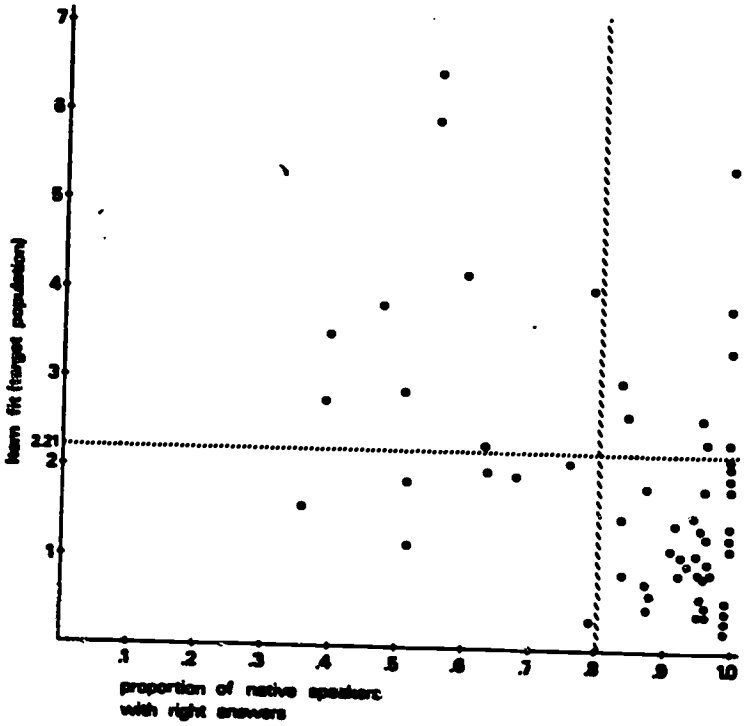


Figure 8B



In figure 8B every dot represents one of the 59 items in the listening comprehension test. The proportion of native speakers who gave right answers on each item (x-axis) is plotted against the measure of fit in Rasch analysis of the responses observed in the target population (y-axis). The critical upper limit of the fit measure at 2.21 is indicated by a horizontal line. The vertical line represents what we take as the lower limit for native speaker response. Most of the items (35 out of 59) fall within both limits. Another 9 items surpass both limits and thus confirm the hypothesis too: misfitting items are too difficult for natives because they measure the wrong trait, they do not test native speaker listening ability. A total of 44 (35+9) items come up to our expectation. Still we are left with 8 items showing misfit that are not too difficult for natives, and 7 items with a low native speaker score apparently fitting in the test: a total of 15 items contradicting the hypothesis. Figure 8A presents the same findings once

against the complete circle represents the whole test of 59 items with the different sections representing the item falling either within or without one or both of our criteria.
Phi-correlation between measure of fit and native speaker score as dichotome variables is .37.

The next step is to delete the 9 items that surpass both limits and do another Rasch analysis on the remaining 50 items. The results are presented in figures 9A and 9B.

After the 9 items have been removed one more item falls within both limits and three more surpass both limits. We have gained four items: a total of 48 now comes up to our expectations. Phi-correlation has moved up to .56.

Figure 9A and 9B: Native speaker response versus measure of fit in Rasch analysis of target population response (50 items)

Figure 9A

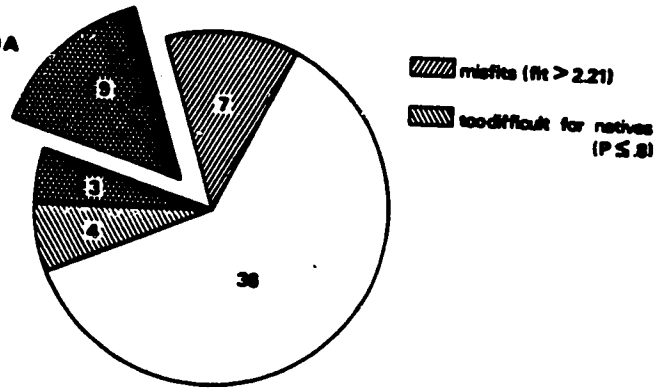
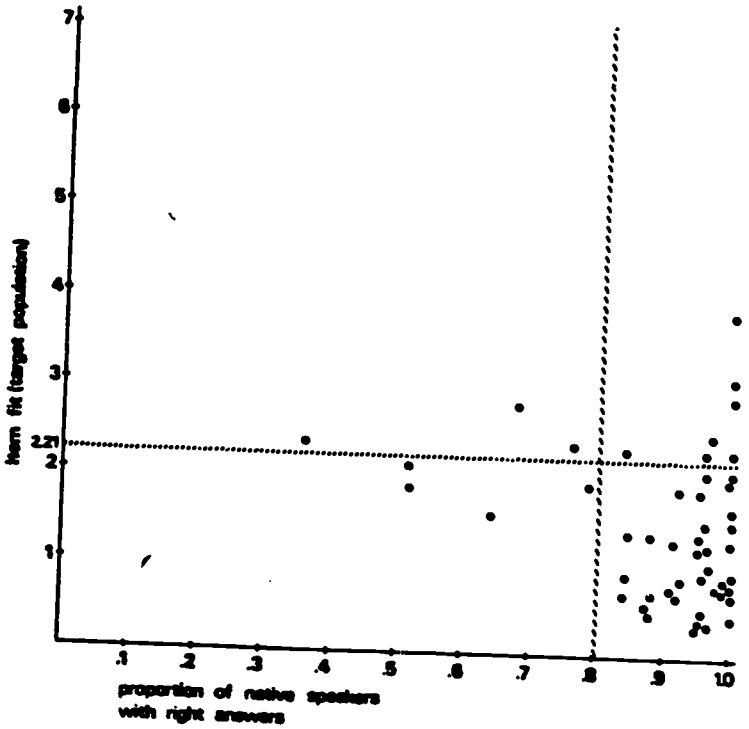


Figure 9B



Another 3 items can be deleted from the test now (figures 10A and 10B) leaving 47 items in the test. Two more items are now within both limits whereas 3 more surpass both limits: a gain of 5 items supporting the hypothesis. Only 6 items are left that do not correspond with our expectations. Phi-correlation is now .77.

Figure 10A and 10B: Native speaker response versus measure of fit in Rasch analysis of target population response (47 items)

Figure 10A

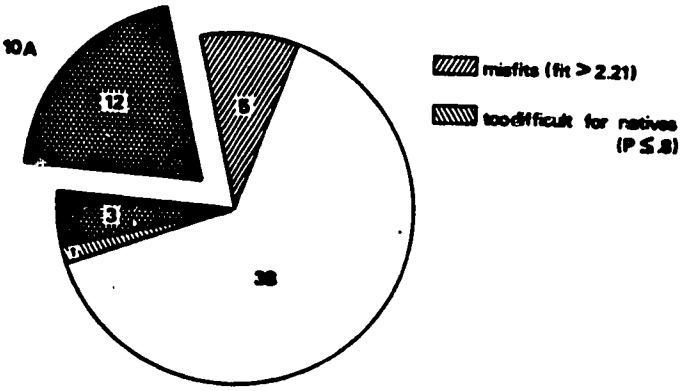
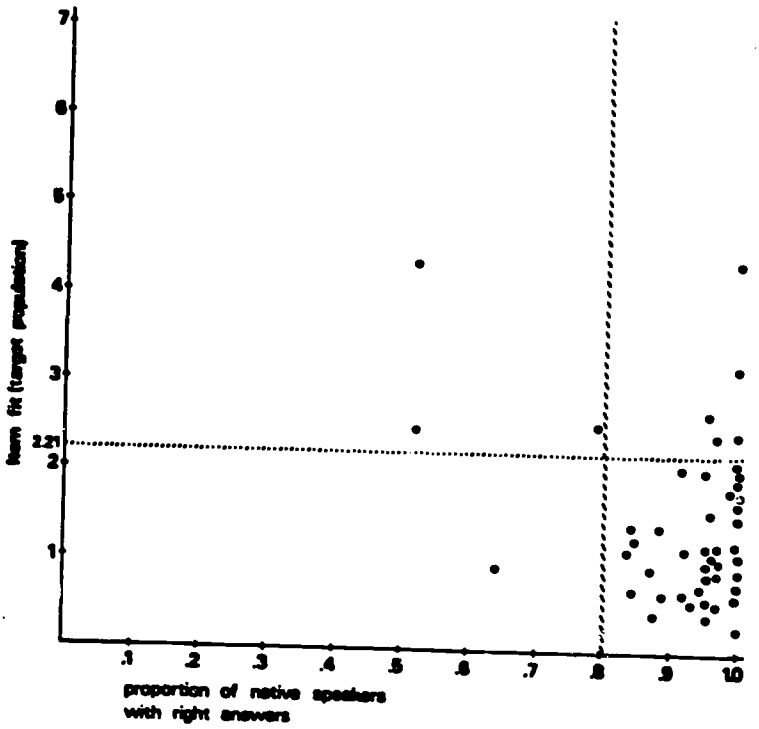


Figure 10B



Deleting the last three items that surpass both limits and reanalyzing the remaining 44 items (figures 11A and 11B) we find one more item to fall within both limits. There are no more items, however, that surpass both limits and the deletion process seems at a dead end. With a total of 54 items in favour of our hypothesis and 5 contradicting it, phi-correlation is now .80. But of the 4 remaining items that apparently misfit in spite of sufficient native speaker score, two items show a negative difference between the mean score of native speakers and that of the target population: the target population did better than native speakers on these two items. This means that the misfit of these items in fact does correspond with our expectation.

Figure 11A and 11B: Native speaker response versus measure of fit in Rasch analysis of target population response (44 items)

Figure 11A

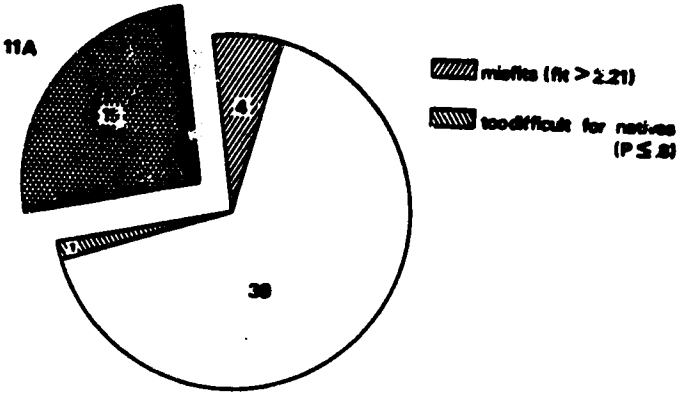
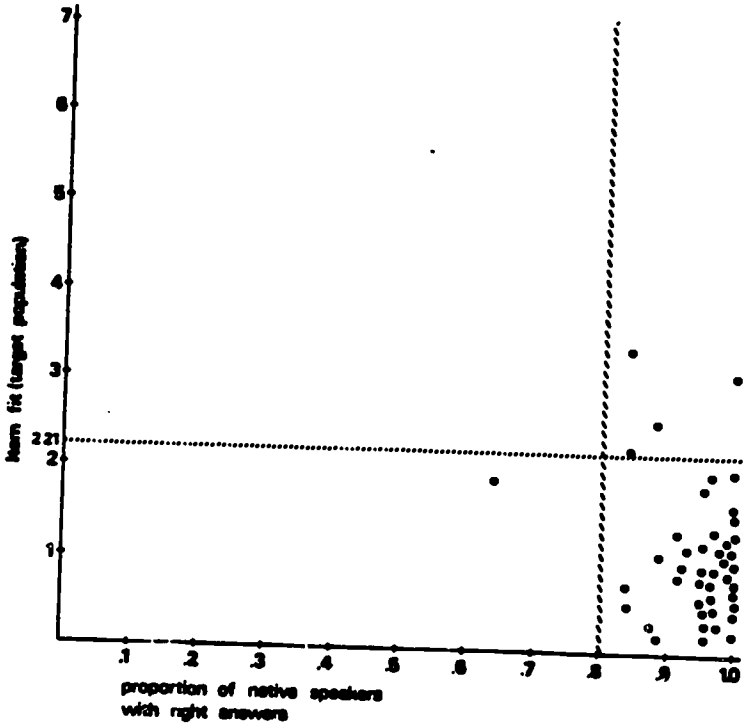


Figure 11B



Figures 12A and 12B show what happens if we delete these two items: one more item falls within both limits and only two out of the initial 59 items refuse to come up to our expectation: one slightly misfitting item easy enough for natives and one item that is too difficult for natives but still fits. The 40 items falling within both limits now cluster in the bottom righthand corner of the graph in figure 12B: these items fit extremely well in the test and present no difficulties for the group of native speakers.

Correspondence between native speaker ability and the latent trait observed in Rasch analysis of the target population response expressed in phi-correlation is .91, and confirms our hypothesis that the latent trait of this test is in fact the ability of native speakers to understand samples of their mother tongue. Thus 40 items out of the original 59 items will make a valid test for measuring listening comprehension of English.

Figure 12A and 12B: Native speaker response versus measure of fit in Rasch analysis of target population response (42 items)

Figure 12A

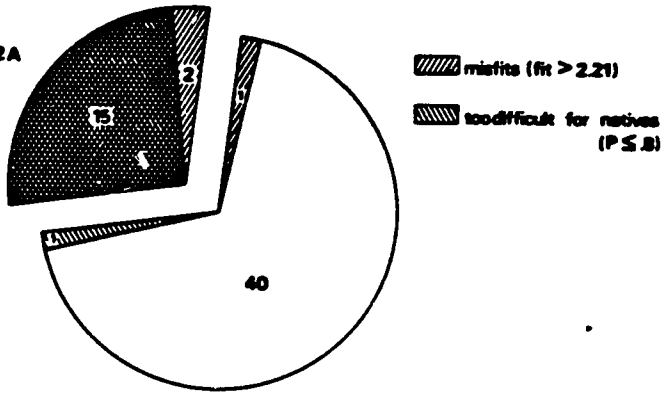
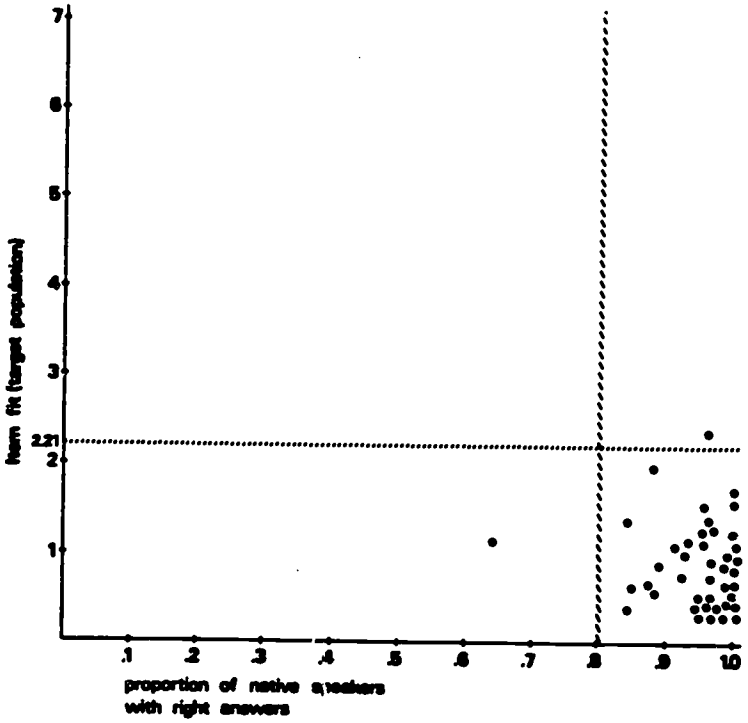


Figure 12B



The remaining 19 items do not measure this ability. Figures 13A and 13B represent the results of a Rasch analysis of these 19 items regarded as a separate test. Figure 13A pictures the items in this test: 15 items surpassing both critical limits and 4 items surpassing only one of these limits: presented as a section from the orginial 59 item test (on the left) and regarded as a separate test (the circle on the right). Figure 13B shows that (apart from one slighty misfitting item) these items will fit together in a test. What this test measures is not clear as yet but anyhow it does not discriminate between able and less able listeners of English of comparable age and educational background.

Figure 13A and 13B: Native speaker response versus measure of fit in Rasch analysis of target population response (19 items)

Figure 13A

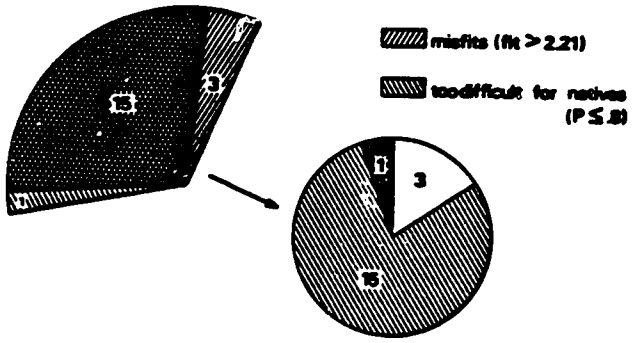
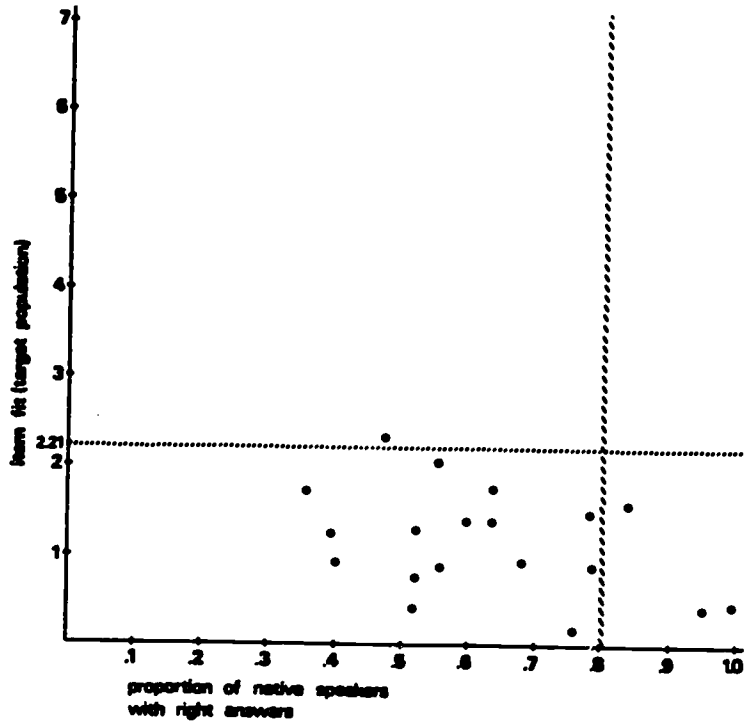


Figure 13B



Examples

I would now like to illustrate the different kinds of items in the test by giving examples of items that do fit in the test and of those that do not fit.

The first example is item number 17.

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...

(Fortescue:) 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.

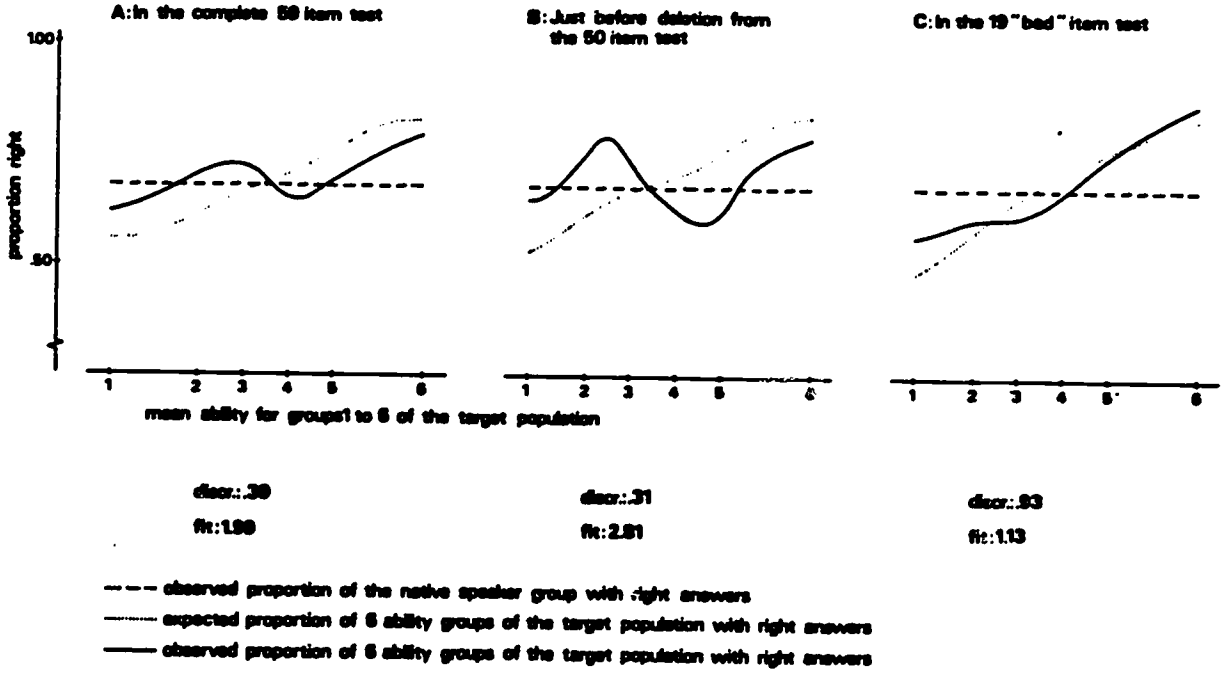
Question booklet:

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

The proportion of the native speaker group that chose the right answer (A) was 1.00 and .86 of the target population chose this answer. All the native speakers found the right answer obvious and for the target population it was an easy item. The item is easy if you understand English. Nevertheless, the item initially showed misfit (3.86) in Rasch analysis of the target population response. Misfit is combined with a high discrimination index of the item (2.35) but disappears after the operation of deleting items showing misfit along with a low native speaker score. Fit measure for item 17 improved and went down from an initial 3.86 to 2.97, 2.47, 2.10 respectively, as more and more items were deleted, in the way I have just described, ending up at 1.73 in the final 40 item 'good' test of figures 12A and 12B.
Figure 14 pictures the item characteristic curves of item 17 in the 59 item and in the 40 item test.

The correct answer is A. A porportion of .68 of the native speaker group and .69 of the target population chose this answer.
Though .69 of the target population indicates that the item is of normal difficulty the native speaker score is far too low. The major part of the target population went 'through the ceiling' scoring higher than the native speakers. What is wrong? Mr. Wilson asks whether people are 'reluctant' to use the new facilities like credit cards etc. Watts's answer seems to be 'yes' because he gives an explanation saying 'credit is an emotive term'. Linking this to 'reluctance' used in the question, the listener can conclude that people do not like the idea of credit. The item is more a test of intellectual gymnastics than a straightforward test of English which explains why it is equally difficult for native speakers and for the target population who have a simular educational background. The item, however, fits in the 59 item test (1.99) but has a very low discrimination index (.39). After deleting the first 9 items in the operation described, fit measure leaps up to 2.82 and the item no longer fits.
Figure 15 gives the item characteristic curves of this item in the 59 item test (figure 15A), after deletion of 9 items (figure 15B) and in the 19 item bad test (figure 15C).

Figure 15: Item characteristic curves for item no 14



Both these examples were true-false items. Item 43 is a modified cloze item with two options:

Item no 43

Tape:

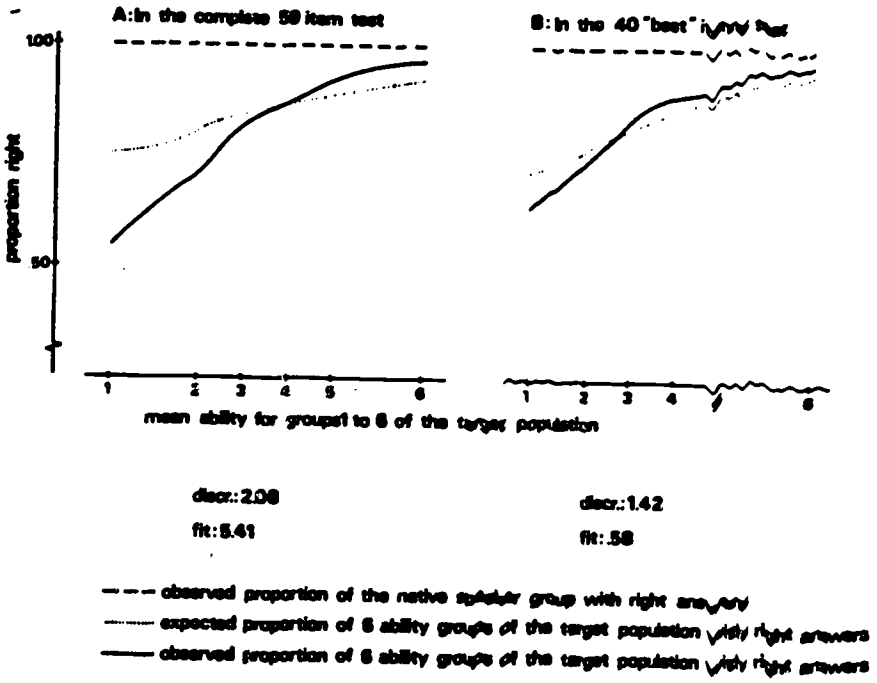
Prince Charles has been speaking about Britain's industrial future. He told delegates at a conference in Bournemouth of the iron and steel trades union that if only the two sides of industry would co-operate more Britain's performance internationally could be —more—

Question booklet:

43 A vastly improved
B very disappointing

Item 43 is a good item: 1.00 of the native speakers and .86 of the target population chose the right answer (A). It is an easy item but not too easy. Nevertheless, the item does not fit (5.41) as discrimination is too high (2.08). In the subsequent stages of deletion the fit measure goes down from the initial 5.41 to 3.85, 3.21, 1.86 to a final .58 in the 40 item test: an extremely good fit! The discrimination index is now 1.42. Figure 16 shows the item characteristic curves in the 99 and in the 40 item test.

Figure 16: Item characteristic curves for item no 43



We continue with item no 44:

Item no 44

Tape:

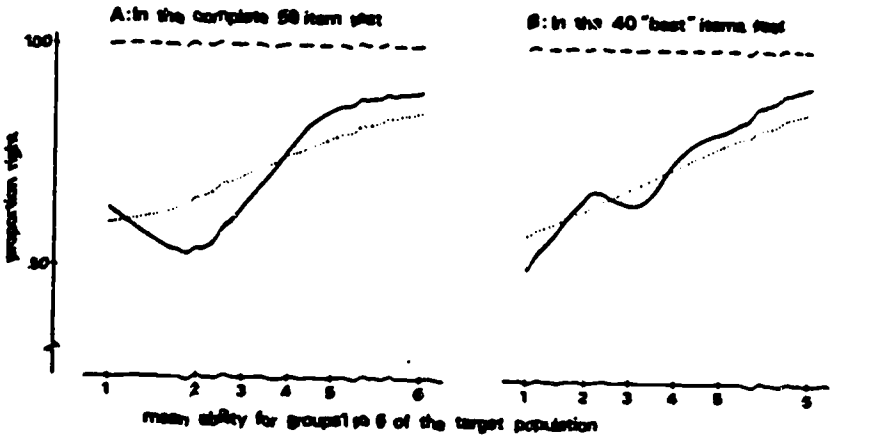
Prince Charles appealed for management and unions to reach a better understanding and went on to express his concern that Britain was ---Ruzz--- its overseas competitors.

Question booklet:

44 A tolerating
B falling behind

Again this item was no problem for the native speakers: 1.00 of this group gave the right answer against .71 of the target population.
Initial misfit of the item in the 59 item test is 3.31 but goes down to 1.73 after the deletion operation.
Figure 17 gives the item characteristic curves.

Figure 17: Item characteristic curves for item no 44



Item no 6 is another example of a trickey true-false item:

Item no 6

Tape:

(Wilson:) Seymour Fortescue, is this evolution to the elimination of money, one which is happening virtually All over the world?

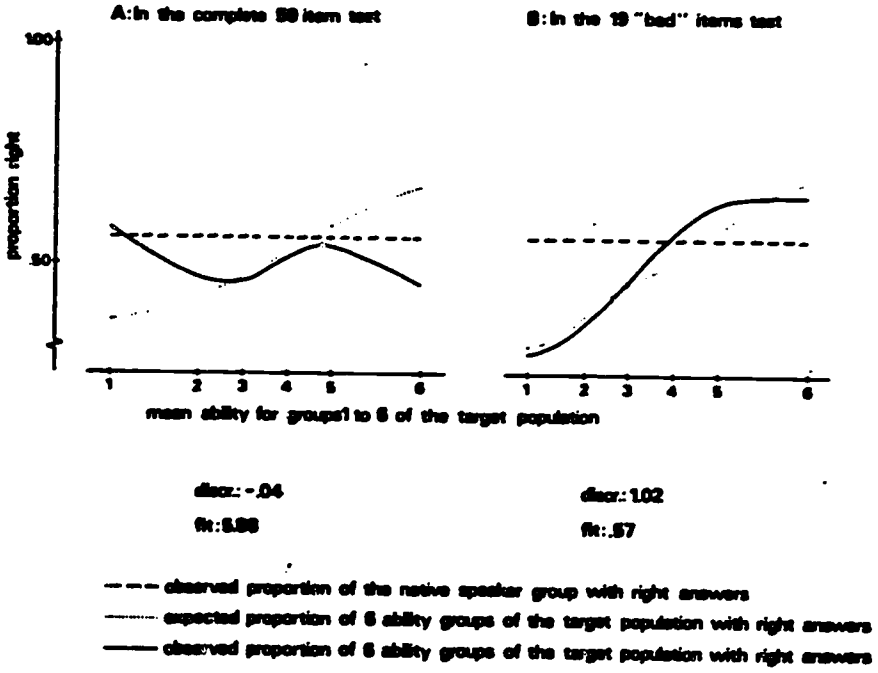
(Fortescue:) Yes I think it is. Perhaps the most interesting place is Japan, because the individual doesn't use cheques in Japan, they use cash and they use pass-books. You can see Japan making the switch from cash to electronic funds transfer without passing through the intermediate stage of cheques.

Question booklet:

6 Japan has already switched from cheques to electronic money transfer.
A True B False

Obviously, native speakers had great trouble in spotting the incorrectness of the statement: only .56 chose answer B. Of the target population a proportion of .52 chose this answer. The question is too difficult; the interpretation process necessary to arrive at the right answer is impossible in the short time allowed by the speed of the ongoing tape. The tone of voice is positive: there is 'elimination of money' in Japan, Japan is ahead in modernizing the payment systems but just because Japan is so far ahead and skipped the stage of cheques altogether the answer should be 'False'. The item shows severe misfit (5.88) and does not discriminate at all according to expectation among the testees in the target population (-.04)! In the 19 item 'bad' test, however, the item shows perfect fit (.57) and normal discrimination (1.02). It seems to me that the 19 item 'bad' test would be a good test of intelligence or alertness but it is not really a proper test of English listening comprehension.
Figure 18 shows the item characteristic curves of this item in the 59 item test and in the 19 item 'bad' test.

Figure 18: Item characteristic curves for item no 6



The last example I wish to discuss is item 45, one of the two with a high native speaker score but an even higher score in the target population, the last two items to be deleted.

Item no 45

Tape:

Uganda's new president, Mr. Godfrey Benaisa, has appealed to the people of the country to co-operate in building a prosperous country that respects the rule of law, the protection of human rights and the establishment
---Buzz---

Question booklet:

45 A of a new state
B of democracy

Native speakers scored sufficiently high on this item: .84, but of the target population .90 chose answer B. One could defend both options, but the word 'democracy' seems to go better with 'respects the rule of law' and 'human rights' and was in fact the original word. Knowledge of political jargon in a revolutionary context and thus general intelligence or knowledge of the world is tested rather than English listening ability. The item fits in the original 59 item test but after several deletion operations it acquires a fit measure of 3.43 and is deleted in its turn. In the 19 item 'bad' test item 45

shows perfect fit.
Figure 19 gives the item characteristic curves.

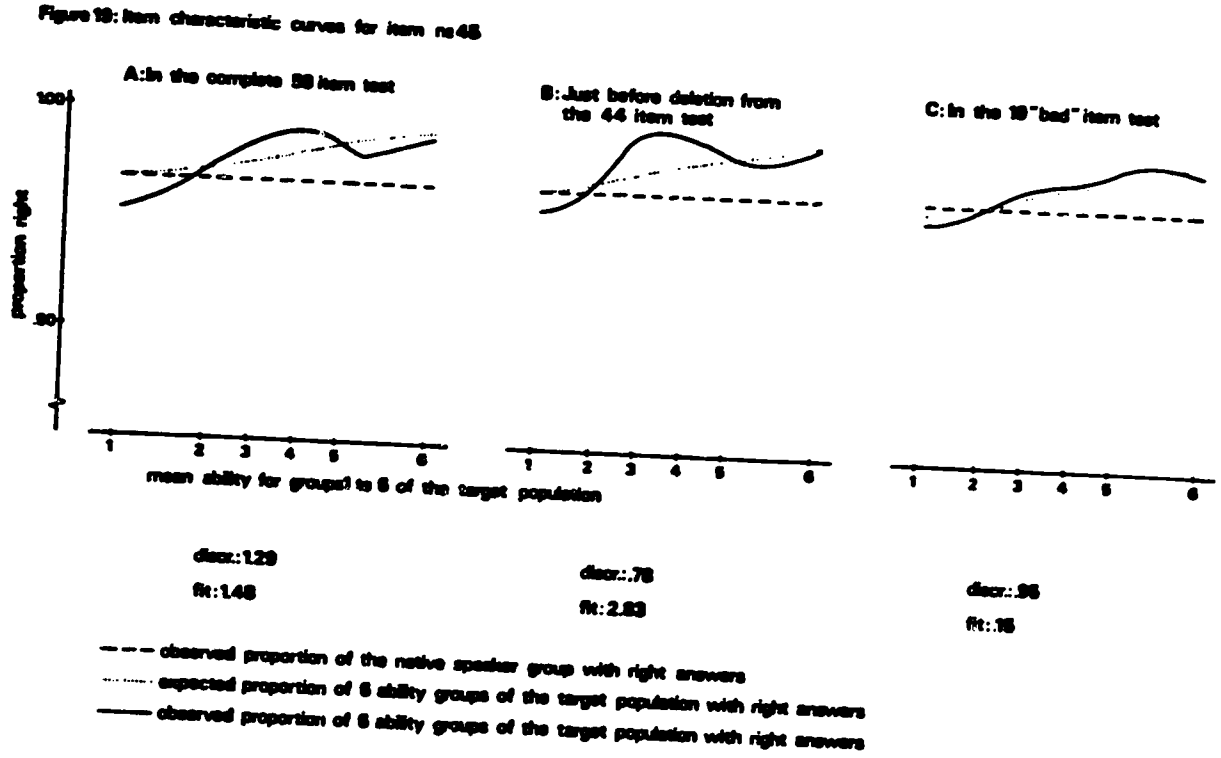
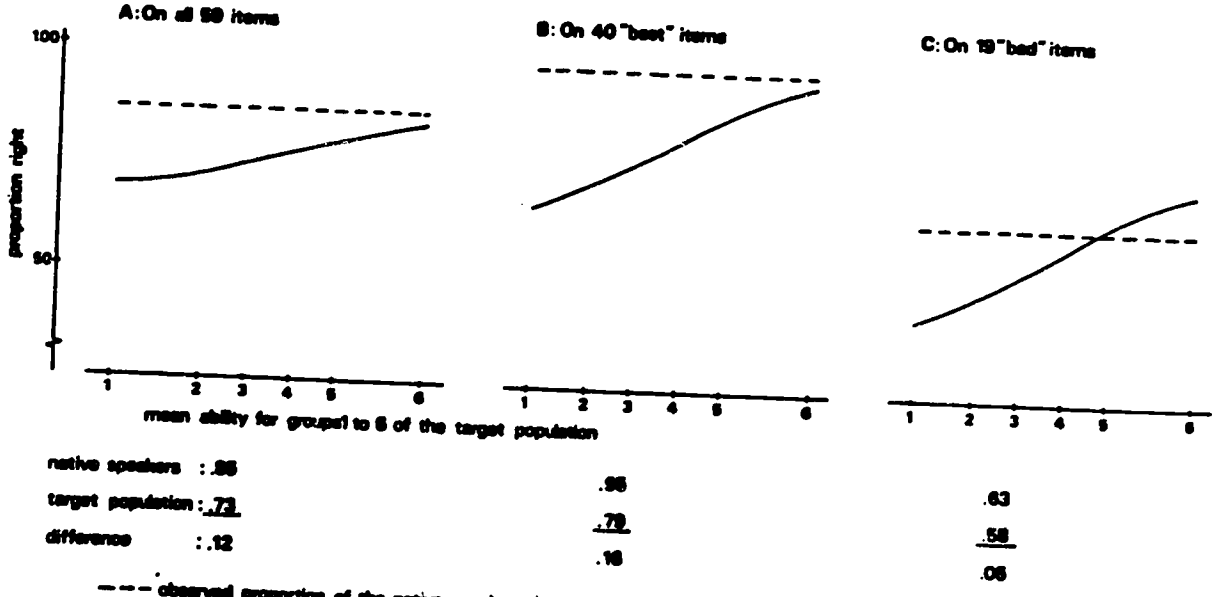


Figure 20 presents the mean of all item characteristic curves and shows the achievements of native speakers and of the six ability groups of the target population on the whole test and on selections of items from the test.



From this picture it is clear that the native speaker group obtained a higher score than the average testee from the target population, higher even than the average testee in the most able group (6) of the target population. The selection entitled '40 best items' creates a larger distance between the native speakers and the average of the target population. Moreover the 40 item selection discriminates better between the different ability groups in the target population. In the 19 'bad' items test discrimination between the ability groups of the target population is about the same but native speakers score only slighty above the average of the target population, native speakers have no advantage in doing this test: it is not a test of English listening comprehension for this population.

One point remains to be made clear. The six ability groups in the subsequent Rasch analyses are redefined for every seperate analysis: the testees are divided into six groups according to their score on the test that is being analyzed. The most able group on the 19 item 'bad' test is different altogether from the most able group on the 40 item 'best' test. To illustrate this, product-moment correlations were calculated between the 19 item 'bad' test and two seperate random selections without replacement of 19 items from the 40 item 'best' test. Selections were made to be of equal length for easy comparison.

Table of product-moment correlations between 3 selections of 19 items from the listening comprehension test (n=575).

	All	A	B	C
All 59 items	x			
A: 19 'bad' items	.40	x		
B: 19 'best' items	.78	.06	x	
C: 19 'best' items	.75	-.01	.47	x

From the correlation indices in the table it is clear that there is no relation between the 19 'bad' items and either of the two selections of 'best' items and testees scoring high on the 'best' items will not necessarily score high on the 'bad' items. The two selections of 'best' items have a positive correlation significant at the .005 level which is fairly high if we consider test length.

Conclusion

The latent trait of the listening comprehension test analyzed here can be identified as the ability native speakers demonstrate in understanding spoken samples from their mother tongue. Two thirds of the test constitute a valid measure for listening comprehension of English as a foreign language. The other third of the test seems more suitable for testing an ability which might be general intelligence or knowledge of the world but does not discriminate between different ability

levels of English listening comprehension.
A more general conclusion could be that Rasch analysis can be of great help in selecting valid items from a try-out version of a test. Provided that the major part of the try-out version does test the ability aimed at, the test constructor, by deleting misfitting items with a low discrimination index in successive Rasch analyses, can make a selection of items that have the highest probability of testing the ability he is aiming at.

References

American Psychological Association (1974) *Standards for educational and psychological tests*. F.B. Davis (chair), American Psychological Association, Washington, D.C.

Cronbach, L.J. and P.E. Meehl (1955) Construct validity in psychological tests, *Psychological Bulletin* 52, 281-302. Reproduced in Noll, V.H., D.P. Scannel and R.P. Noll (eds) *Introductory readings in educational measurement*, Houghton Mifflin, Boston 1972, 90-121.

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

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

Vollmer, H.J. and F. Sang (forthcoming) Competing hypotheses about second language ability: a plea for caution. To appear in J.W. Oller (ed) *Issues in language testing research*, Newbury House, 9 Rowley, Mass.

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.