A view of heather, fir coppices and unipers, Anholt, Denmark

This article was first published in *Language Testing* vol. 11. # 1. 65-81.
POSTSCRIPT The use of 'Sprogtest' was discontinued when we were informed that better methods had been developed. This information has turned out not to be correct, so the article may perhaps be of help to others.

'SPROGTEST': A SMART TEST , or HOW TO DEVELOP A RELIABLE AND ANONYMOUS EFL READING TEST)

Cay Dollerup, Esther Glahn and Carsten Rosenberg Hansen
Copenhagen, Denmark

This article deals with a Danish English-language reading proficiency test. It is dubbed 'Sprogtest' which means 'language test'. Since English textbooks are used extensively in Denmark, the test is offered to freshman students in order to diagnose weaknesses which may impede the undergraduates' academic careers. The test has to be anonymous and convincing to those who use it.
In order to facilitate the immediate assessment of what parts can be transferred and used in other language areas, the article discusses the test construction, development and improvement in detail.

I 'Sprogtest': background

Denmark is a small country with only 5.1 million inhabitants.[1] Therefore there is no large-scale readership for Danish textbooks at Danish universities, and so English-language textbooks are used extensively.

Although a 'numerus clausus' (based on the average college grade) was introduced in 1977, there is no formal entrance examination at university level. Since the 'numerus clausus' disregards specific grades, there is no control to ensure that the students' reading comprehension of English enables them to understand the English textbooks.

When textbooks are not understood, teachers are therefore not in a position to determine whether the reason is (a) a failure on the students' part to understand written English; (b) scholastic lack of aptitude; or (c) that the textbooks are too difficult.

Faced with this problem, university teachers turned to us in the early 1970s for an 'objective assessment' of reading comprehension of English among Danish freshmen undergraduates. It was obvious that this could be done only by means of a test which fulfilled the following requirements:

…// 66 …

1) It must be reliable and 'objective', insofar as every participant's performance was assessed by the same criteria and in precisely the same way, i.e., there must be no teacher subjectivity.

2) It must test reading comprehension in English in general, in order to be useful at all Danish seats of higher education.

3) The test must not comprise specialist terminology, partly because this would limit its applicability and partly because the correct updated terminology would be used in classes anyway.

4) Teachers must be able to assess the overall proficiency of classes in order to decide if the textbooks might be too difficult.[2]

5) Testing and feedback had to be anonymous in order to avoid any suspicion that it was a camouflaged entrance examination test - especially so since the 'numerus clausus' is a sensitive issue.

6) Students must be informed individually about deficient reading comprehension before it leads to poor academic performance or failure. The test should be used only at the beginning of the students' academic careers, namely within the first two months of their university studies.

II Constructing 'Sprogtest'

Our approach was pragmatic in the selection of texts: in order to cast as wide a net as possible in reading comprehension, we chose to use three types of texts, namely popular science articles, newspaper articles and fictional texts.

Within each type we started with a dozen potentially useful texts. They were authentic and either complete or rounded excerpts of 250-400 words. Thirty readers then underlined unfamiliar words. Low-frequency words (below the 5000-7000 frequency bands according to Thorndike-Lorge) unknown to three readers or more were usually replaced with more frequent synonyms.[3] It was our assumption, then (as teachers), and it is our conviction now (as researchers - see Dollerup, Glahn and Rosenberg Hansen, 1989) that readers who have trouble with individual words simply have vocabularies of less than 5000-7000 words.

…// 67…

Subsequently we made multiple-choice questions for the 12 texts. The texts were tested out with readers. After each testing round we replaced distractors scoring less than 10%. In this way, the number of texts was successively reduced by a process of the 'survival of the fittest'. The 'fittest' turned out to be three texts all of which had the multiple-choice questions in the run-on texts, not at the end. [4] In order to avoid confusion, the texts have retained their original code numbers through all our work, so the 'survivors' were '230' (a newspaper article), '330' (a popular science text) and '420' (a passage from a short story). They were put together into one test comprising a total of 45 questions (ranging from 13 to 17 per text). After a final check under field conditions we decided to launch the test nationwide in 1976 and, thanks to grants, it has been possible to offer it free of charge since then to freshmen (Dollerup and Dinsen,1978; Dollerup, Glahn and Rosenberg Hansen, 1978; 1980).

1 Safeguarding anonymity

It will be recalled that anonymity was of prime importance to us. Furthermore, it must also be obvious to participants that they cannot be identified by others than themselves, otherwise they would not use the test.

For this reason we developed a strict procedure for handling the test.[5] The test is advertised in university newsletters. Interested teachers contact us, and our assistant prepares the test. On the first page of the test, she numerically enters the year, the month, the institution and the field of study (in case we should wish to conduct longitudinal studies later on). This is followed by a'group number' and a three-digit 'personal number'. This number is filled in consecutively within each group. There is also a covering sheet instructing participants that the test will last about one hour; that they should not use dictionaries; that it is important that all questions are answered, even if they have to resort to guesswork; and that they must copy the 'personal number' from the front page of the test in order to be able to identify themselves in our feedback.

This material is sent to the teachers.

… // 68 …

2 The test

The opening passage containing the first six questions in Sprogtest is shown in Appendix 1.

3 Feedback

The completed tests are sent to us and the information is fed into the computer, group by group. The computer prints out lists with feedback on individual performances in each group. These lists are forwarded to the participants with covering letters. The original tests are destroyed.

4 Feedback,1976-89

Until *1989*, the feedback to the students would consist of four parts. At their 'personal number' was (1) a raw score for correct answers to each text. A weighted score (2) (taking into account the gravity of errors) placed the student in (3) proficiency categories A, B, C or D.[6] A feedback letter (4) gave concrete advice to poor performers about procedures for quickly improving their reading comprehension. It also gave students information about the specific character of each text, in case they had problems with only one.

From our point of view, the advantage of Sprogtest was primarily that teachers could use it for exhorting students who had problems with their English to do something to improve their reading competence. The 'objectivity' of a computer printout enhanced this effect.

In constructing Sprogtest and in our use of it, it was important not to make exaggerated claims about it. It functioned well, for it was used every year by some *100-600* students and was frequently requested repeatedly by the same institutions.

III Support work

By means of Sprogtest it is possible to conduct discreet studies of differences in proficiency between faculties (e.g., engineering vs. the humanities) but, so far, no such difference has been found. Our only result from comparisons was unexpected. One institution used the test the year before and the year after the introduction of the 'numerus clausus'. It showed a jump in reading proficiency of more than 5%. A journalist got wind of it and wrote an article justifying the 'numerus clausus'. We asked the teachers involved if they had noticed an improvement in student performance. … // 69 …They had, but had attributed it to the introduction of a new textbook.

IV Updating the test

By 1989 it was clear that if Sprogtest was to go on carrying conviction it must be overhauled. This was mostly due to improvements in printing techniques which made our original material look dated. But we also had tangible problems: the programme for assessing student performance had been transferred from one computer system to another, often by new assistants who had added something of their own: the input side had been clogged up. On the other hand, the test base now included information from 1616 readers.

This supplied us with a numerical and empirical basis we had not had previously. The revision came to comprise retyping, an analysis and an improved feedback. We are concerned only with the latter two in this context.

The numerical information supplied us with precise information about how the questions and distractors had fared. This is shown in Appendix 2, a sample from the first three questions of the test (text '230', which was also cited in Appendix 1). It will appear that most distractors functioned well, although there were a few (e.g., distractor 3 in question 1) which now failed to attract 10% or more of the readers.

The numerical information served as a basis for assessing the relevance of the findings of an introspection study conducted ten years before. We therefore returned to this introspection study.

V The introspection study

The introspection study (1979) had been undertaken in order to procure concrete information about

1) problems Danish readers had with Sprogtest;
2) features in the texts which caused problems; and
3) strategies and mechanisms used by readers for choosing one specific alternative rather than another one.

We have often used introspection (think-aloud protocols) in Denmark in reader response studies (e.g., Dollerup, 1971) and for studying non-native speakers' production and communication strategies (Glahn, 1980).[7]
… // 70 …

There were 28 participants in the 'Sprogtest' introspection study. Seven were university students majoring in English (three men and four women, aged 23-26; numbered 1 to 7 below) and 21 students from 2.g. at a'gymnasium' (i.e., college, *lycee, Hochschule)* (four men and 17 women aged 17-19; subsequently numbered 8 to 28).[8] The interviews were undertaken by Cay Dollerup: care was taken to establish a relaxed atmosphere; the participants were told to act as they normally would. The readers were asked to report during the reading of the texts and to explain why they picked one alternative rather than the other two. In order to prompt these explanations, they were often asked, 'How did you identify that?' in a tone implying admiration and approval. In general, Danish readers found it easy to report about their response to the English texts in Danish.

After the reading there were supplementary questions. The readers' reports were taken down as notes and taped for control; and we produced a fifty-page hand-written report which was used for the revision of 'Sprogtest'.

1 The differences between the university and the college students

It goes without saying that the introspection study did not elicit information about all alternatives.

Some 11% of the options chosen by the students and 30% of those in the college group were not explained. But the study still provides about 950 explanatory comments for the 45 questions (and 135 options). Being majors in English, the university students were good readers who mostly supplied us with information about strategies leading to correct answers. Conversely, the average performance of the college students (with an average of 22 wrong answers out of 45) is poorer than the average in Sprogtest. In all likelihood the introspection method made for more errors than normal reading. Yet the erroneous answers showed overall patterns and correspondences which permit us to assume that they really shed light on the selection of wrong alternatives in Sprogtest.

2 The contents of the speeches

There were considerable differences in readers' reports. …// 71 … Most readers, for instance, only told us why they chose one alternative at a question, and it was rare for them to discuss two options. For example, '[230] 21.2. That sounds most correct to me. I was in some doubt about *obstructed,* but I come down on the side of *neglected.'*[9]

Reader 4 tried to imagine the errors others would commit: '[230] 21.1. Both options 2 and 3 are found in clichés in Danish, so one would hardly choose them', '22.2. *Closed* is an unfamiliar word. And *safety work* was mentioned in the first line', '23.3. This is because we hear of *voluntary safety work.'*

And Reader 9, who performed well, often resorted to the empty catch-phrase: 'This sounds right to me.' There were surprisingly few attempts on the readers' part to go back and correct answers they must later have realized were wrong. It happened only in five cases, i.e., in less than half a per cent.

The majority of the answers were ambiguous and complex. On the one hand, answers from different readers might be conflicting: in the same fashion that different arguments might be used to identify the same alternative in the multiple-choice questions, near-identical arguments might prompt readers to pick different options. Furthermore, many answers contained several items of information at varying levels. Thus, for instance, the statement, '33.2. This gives the best sense'. *Interviewer:* 'How's that?' 'When I read the context - there are words I do not understand. And by means of a method of exclusion and by pronouncing it in English, as well as a translation into Danish, that's the alternative which sounds best', contains information on (a) an attempt to use a strategy for making sense which flounders on a failure to understand all words; (b) a poorly illuminated method of exclusion; and (c) a check on euphony in English, and finally a translation into Danish.

The complexity of the answers means that the strategies we could discern in the answers are questionable insofar as they do not necessarily reflect on the total reading process because the study was limited to these three texts. On the other hand, it is in keeping with our overall attitude that in order to illuminate this particular reading situation, we must use the information relating to the test rather than extraneous sources. At the same time, the complexity of the results imply that we must show discretion in the analysis.

Allowing for the multifacetness in the readers' answers, it still seemed as if we could

identify some strategies which cropped up tolerably often with different readers independent of whether they led to the correct option or the two wrong alternatives. …// 72 … They could be defined albeit with caution. A list of strategies generated from the readers looks as follows.

(a) Guesswork: Guesswork was represented by shots at the multiple-choice questions, without any attempt to explain and justify it. It will be recalled that readers of Sprogtest are admonished to complete all questions even if they have to guess, and therefore participants in the introspection study were similarly asked to guess alternatives in case they could not identify the 'right' one. The underlying idea was to find out if 'guesses' would have some kind of implicit rationale which meant that it tended to lead to either right or wrong answers. However, it did not: the answers were arbitrary, so guesses in Sprogtest hardly bias answers towards specific options.

(b) Euphony: '24.1. I have absolutely no idea, but it sounds best.' '24.3. I don't know what *precautions* mean. But I have a feeling that *in the light of* sounds best here.' The strategy is not quite clear, but the phrasing is indicative of a general disorientation, where an attempt to hear how the options sound (if only mentally) becomes the main prop.

Like guesses, euphony had no consistent pattern and it led to both right and wrong choices.

(c) Decoding individual words: Unfamiliar words led to vacillation, to guesses and to attempts to grasp their meaning. In the course of the test solving, readers used different methods:

1) They resorted to their knowledge of other languages (such as Latin).
2) They used their knowledge of Danish, for instance, of *false friends* or just some superficial similarity.
3) Readers mistook one word in English for another: Reader 28, for instance, mistook *ensured* for *assured* in '[220] 22.3. *Closed* since the management had to *assure* that mines remained closed during the strike.'
4) In some cases, specific (strong) words affected the choice: '[420] *A man whose respectful manner ... showed that his position was one of /1. authority /2. equality /3. dependence.* It cannot be 2, for *manners* exclude *equality*. I opt for 1 because of *manners*. They show he has some standing.'
5) The context would often furnish readers with a clue - right or wrong - which they would then use. The way some readers tackled the word *pit* in the first text serves as an illustration: '[230] 22.3. *This time co-operation between miners and management ensured that the pits were closed:* I assume the miners have some *offices* or some *shop stewards*, that's what *pits* must be.' …// 73 … *'The men went down to see what was happening, and /1. where necessary /2. not /3. to discuss whether to take action.* If *pits* are *shop stewards* they must discuss it.'

It is surprising that readers were equally prone to accept the unfamiliar words and phrases as correct alternatives, and to discard them as incorrect.

(d) Syntax: Reference to syntax would normally lead to the right alternative: *'[230] 25.2.* The next line must refer to *them.'*

(e) Combinations of textual information: This strategy might involve several bits of erroneous information as in the following two readings of *330: 'Floods in India and Bangladesh. A prolonged dry period in Africa. These widely reported /1. floods /2. phenomena /3. underdeveloped countries all have something in common.* I don't really understand this, but since both India and Bangladesh are underdeveloped countries, it must be *3.'* And *'3.* The reason is that they talk about floods in India and Bangladesh.'

(f) Anticipation: Anticipation, defined as an attempt to predict what will follow in the text, is based on the combination of textual information. This argument was found exclusively in the literary text, which indicates that this is a strategy which is prominent in the reading of narrative fiction: *'[420. On the 3rd of June, 1890, a gentleman, who gave his name as Monsieur Louis Caratal, desired /1. to send a letter to /2. an interview with /3. to follow in the footsteps of Mr James Bland.]* 1. It is because he tries to find out the background of the story.'

(g) Common sense: This is best defined as 'what is sensible'. And it was not a waterproof method. This is borne out by two readings which attempt to set up common-sensical explanations of the passage *'miners worked to reduce water levels after /1. a spontaneous outbreak of fire /2. a major pump breakdown /3. a collapse of roof supports.* 1. It may be that they had those fires and then there was too much water in the mines.' And: 'There can't possibly be a fire in a mine. It must be the roof caving in.'

(h) Background knowledge: Background knowledge was of little avail to readers. More often than not it led to wrong alternatives: *'(Floods in India and dry period on Africa] were caused by 1. Nature/2. agriculture /3. industrial pollution,* 2, because agriculture is the principal industry in most developing countries.' …// 74 …

We believe this is the very reason why these three texts survived the rigorous selection in the construction of the test: readers cannot pick the right alternative out of context, just by using common sense. This is a weakness in many multiple-choice tests.

(i) Translation into Danish: This method is often used in English-language teaching at *Hochschule/college-level* in Denmark to check comprehension. When employed, it usually led to correct solutions: '24.3. I picked 3 because it means *given the experiences from the previous strike.'*

This method could hardly be registered in 'Sprogtest' itself. Still, our experience with teaching foreign languages means that in the feedback letter offering advice to poor performers, students have always been told to use this strategy, so it was reassuring to find empirical confirmation of our pedagogical counsel.[10]

There are undeniably questionable overlappings in the above categorization. The categories are, of course, also more obvious to us, for the simple reason that they derive from Danes and are set up by other Danes. Danish teachers of English as a foreign language will be familiar with 'euphony' as a strategy among tolerably advanced students whereas it must seem far-fetched to native speakers of English. Similarly, the differentiation between 'common sense' and 'background knowledge' will obviously vary from country to country: readers in countries with coal mining (for instance the UK) will be more familiar with the fact that mines must be maintained in case of strikes than people in countries without mines: having no national mining industry worth the name, Danes' knowledge of coal mining is superficial - at best.

The harbour, Anholt, Denmark

3 Mainline vs. fragmented reading

In the introspection study there were two markedly different ways of reading. The first one was found primarily with the university students who would quickly grasp the main point of the text, for instance, that 230 would deal with 'Something positive, it has to do with safety'. And: 'Production must be resumed after the strike.' The second was a fragmented decoding found with the college group: comprehension units seemed limited to one sentence, and evocative words might colour the understanding of the whole text. …// 75 …

Thus, for instance, the word *strike* cast the first text in a negative light (see Appendix 1).

4 The breakdown

In view of their ambiguity, complexity and variation, the data from the introspection study were hard to

handle. In order to get an overview, we edited the readers' statements for each question and each option. To give an impression, we quote the summaries from the first questions of the test (cited in Appendix 1):

21.1 (the correct alternative). It is positive// Production must be resumed after the strike// The best alternative in the context// A guess// Sounds best// Voluntary// The other words are unfamiliar. 21.2 It fits with *work/ Obstructed* fits in with a strike// A guess. '
21.3 *Coal mines could resume* fits with *neglect* (this reader refers to the Danish word 'negligeret' which means 'ignored')// They refuse to work during a strike// Because of what follows// Fits best// A guess// It sounds best// Other: 2 and 3 would be clichés// 2 and 3 would be contradictory// 2 does not
connect// 3 is an unknown word// 2 and 3 are unknown to me// Vacillation between 2 and 3// 1 is out of the question// 2 is bad// 2 and 3 are out of the question.

VI Discussion

The breakdown of components and strategies in the reading and test-solving process in the introspection study served for an interpretation of the readers' marks in Sprogtest. It is eye catching that in authentic texts like these, there is rarely one single reason why a given alternative is chosen by a reader.

We have only 45 questions, that is a total of 135 alternatives to go by, and that is far too few to cover all features in reading. Specific sources of error may therefore not appear in readings at all, or only with one or two alternatives - and perhaps with specific readers. Accordingly, we cannot cover all components of reading comprehension of English in the test and in our feedback.

Yet we may assume that if a feature makes the reader answer incorrectly twice, we are in all likelihood in the process of identifying a weakness with this particular reader. The more so, if it happens the third time, and so on.

On the other hand, the categories generated by the readers' answers might not be relevant in terms of the reading test itself. In addition, they may not be generated by many questions.

It was noted how, although present, guesswork, euphony (categories 1 and 2), most vocabulary problems (category 3a-c and e) and what we have defined as 'common sense' did not seem to bias answers towards specific alternatives, but appeared to be distributed in arbitrary fashion at all three options. …// 76 … Accordingly we are sure that they are factors in the reading and test-solution processes, but we cannot put this knowledge to any use.

Anticipation (category 6) was limited to a few questions in the fictional text and thus outside the province of the main objective of Sprogtest, namely to cater for all freshmen at national level.

1 The five categories

Conversely, our breakdown revealed that there were sufficiently many questions where, with due caution, we could use the other categories, in part in their entirety or in combination with the 'mainline vs. fragmented reading', to set up some categories of errors that we dare identify to readers in the Sprogtest feedback:

1) There are six distractors that may be chosen out of ignorance of specific words. Here a total of three ticks lead to the feedback: 'Your English vocabulary is small.' This feedback connects with category 3 above.
2) There are 16 distractors which may be affected by a bias introduced with one word, e.g., *strike* or *manners* (both discussed above). Readers who choose 50% of these options are told that 'You attribute too much importance to special words and phrases'. Feedback of this type is based on category 3, specifically subgroup d and, to some extent, on category 5.
3) There are nine distractors which may be chosen because of incomplete comprehension of the syntax, e.g., 220: 24.1. Readers who pick six of them are told that 'You have trouble understanding syntax'.[11]
4) There are 14 distractors which may be marked because readers rely too heavily on their background knowledge. They include the alternatives in 220:21.2, 22.3, 23.1 and 24.1 (see Appendix 1). Provided readers mark at 50% of these they are informed that 'You rely too much on your background knowledge and you do not take into account details and modifications in the texts'. This is based on category 8.
5) Finally, there are 13 distractors which may be ticked by readers sticking to the immediate context of the passages; in this case seven erroneous marks lead to the warning: 'Your reading is not fluent. You read sentence by sentence instead of getting an overall view of the contents of the text.' …// 77 … This, then, is used because we assume that we have uncovered a problem with readers who use what we termed 'fragmented reading'.

We do not stick to the same percentage of erroneous ticks to make our comment, for a fixed percentage would imply that difficulties are found in equal measure in all texts, and this they are not.

VII The new feedback

In our revision of the feedback we felt no need to change the raw scores and the weighted scores. However, instead of A, B, C and D, we now say 'excellent', 'good', 'moderately good' and 'poor'. In addition, we have added the information about individual weaknesses so that they come at the end of the feedback to individual participants. The feedback is shown in Appendix 3. The feedback thus combines an overall assessment with individualized information to weak readers.

VIII Conclusion

We have described the construction and updating of a test of English reading comprehension for Danes. In order to be used, it must be anonymous, which has demanded careful implementation. And the test must also be convincing to users. These aims have been achieved.

The test uses (largely) unedited, run-on texts in order to reflect faithfully students' reading comprehension of English. The dimension of 'language for special purposes' is deliberately disre-

garded since the texts should be understandable to 'educated general readers of English as a foreign language'. The procedures in the construction of the test have been pragmatic and based on information which has been derived from the texts used, notably in terms of readers' reactions and comments on the texts during reading. These comments have not been clear and unambiguous but complex and fuzzy. There is no obvious or comprehensive framework which could be used for the classification of the reading strategies, and we have therefore had to set up our own.[12]

The use of introspection studies for improving feedback in the test itself is an interface between qualitative evaluation and quantitative assessment. …// 78 … The introspection studies cannot realistically be made to comprise enough readers for quantitative explorations of central issues in the reading test: time, trite repetitiveness in the readers' statements, and lack of resources make this an impossible approach. Yet the quantitative approach represented by Sprogtest is a reliable instrument for measuring but, until we started cautiously to apply results from qualitative studies, we had to rely mostly on intuition in our advice to students, and we had little concrete evidence of why the test worked.

A purely qualitative approach to the data from the introspection studies allows for an interpretation of these data in another light, namely that of teaching English reading comprehension for foreigners. It then seems that reading comprehension in students can be improved substantially by (a) learning new words systematically (to be blunt, by rote), (b) regular drill of grammar and punctuation reflecting on syntactical relation, (c) regular textual references (to avoid reference from clichés and noise from 'general background knowledge'), as well as (d) training in reading larger segments at a time.

In the feedback in the test itself, we cannot go to such qualitative extremes. On the contrary, many of the strategies that were unearthed in the qualitative analysis did not seem to be easily put in quantifiable terms, mostly because their effect on the actual test solving and hence on what is gaugeable (the ticks) is arbitrary and unsystematic. It is only the strategies which the qualitative study shows have some kind of systematics and therefore point towards specific alternatives which can be employed for the quantitatively based feedback.

In the updated test, it is probable that, more often than not, we spring not only a'poor' on students whose English is not good but also specific information about their individual weak spots. We cannot really reach the good students and identify their weaknesses, even though this would have been more prestigious and sophisticated. But in our social context that is not the issue: the point is to warn students, whose academic careers may be jeopardized because of deficient knowledge of English, discreetly, and in a cheap and fast way. This is the only way they will stand a chance of improving theirEnglish.

In a larger perspective, the test is interesting in two ways. First, it shows that reliable testing of reading proficiency in a foreign language can be carried out on a large scale. Secondly, the existence of Sprogtest shows that it is possible to operate a test nearly 'untouched by human hands', as it were, and to safeguard the anonymity of its users. …// 79

IX. References

Block, E.L. 1992: See how they read: comprehension monitoring of Ll and L2 readers. *TESOL Quarterly 26,* 319-43.

Dollerup, C. 1971: On reading short stories. *Journal of Reading* 14, 445-54.

Dollerup, C. and **Dinsen, E.** 1978: Comprehension of written English texts among Danish freshman students. *Third European Congress on Information Systems and Networks: Overcoming the Language Barrier.* Munich: Verlag Dokumentation 1, 63-83.

Dollerup, C., Glahn, E. and **Rosenberg Hansen, C.** 1978: On the construction of a test in reading comprehension. In Gregersen, *K. et al.*, editors, *Papers from the Fourth Scandinavian Conference on Linguistics.* Odense: Odense University Press, 79-86.

1980: Some errors in reading comprehension. In von Faber, H., editor, *Leseverstehen im Fremdsprachenunterricht.* Munich: Goethe Institut, 238-56.

1989: Vocabularies in the reading process. In Nation, P. and Carter, R., editors, *Vocabulary Acquisition: Aila Review 6,* 21-33.

Evans, E.E. 1988: Advanced ESL reading: language competence revisited. *System* 16, 337-47.

Færch, C. and **Kasper,** G., editors, 1987a: *Introspection in second language research.* Clevedon and Philadelphia PA: Multilingual Matters. 1987b: From product to process - introspective methods in second language research. In Færch, C. and Kasper, G., editors, *Introspection in second language research.* Clevedon and Philadelphia, PA: Multilingual Matters, 5-23.

Glahn, E. 1980: Introspection as a method of elicitation in interlanguage studies. *Interlanguage Studies Bulletin 5,* 119-28.

Grotjahn, R. 1987: On the methodological basis of introspective methods. In Færch, C. and Kasper, G., editors, *Introspection in second language research.* Clevedon and Philadelphia, PA: Multilingual Matters, 54-81.

NOTES

1. We thank the Danish Research Council for the Humanities for funding the development of the test. Over the years the Tuborg Foundation, Fabrikant Otto Johannes Bruuns Fond and the 'Tips/Lotto' Foundation have supported the development and the maintenance of the test.

2. As far as we can ascertain, no teacher requisitioning the test has ever used it for this purpose, although we always mention it in our covering letters.

3. It will be noted that we used the students' marks as the point of departure, not the frequency of the words. The reasons were many: if we had automatically applied ThorndikeLorge (or any other frequency list) for our substitutions, our editing would have disregarded the importance of such factors as similarities between Danish and English, English loan-words in Danish, the changed importance (and hence frequency) of specific words between the time the count was made and the test was constructed, and so on.

4. Each of the 12 texts used at the start had multiple-choice questions, and each text consistently had the same type of multiple-choice questions all through. About half the tests had the questions at the end. We also used multiple-choice questions with two and four options. However, they did not work well in the testing rounds and were therefore discarded.

5. In order to make sure that the procedure is always followed, we made a manual of about 20 pages containing a step-by-step instruction in what to do and which papers to place where, forward to people, etc. This self-instructive manual has been regularly updated and is handed over to new assistants.

6. 'D' is the score a student will attain by pure guesswork, i.e., 33% and less.

7. Discussions of the methodological aspects are found in, e.g., Grotjahn (1987) and in Færch and Kasper

(1987b).

8. We assumed - and have no reason to doubt it - that there was no appreciable gender difference at this level. The high number of women merely reflects the fact that we interviewed students from the modern language branch of the 'gymnasium' which was vastly more popular with women than men.

9. The following signs are used in the quotations from the protocols: *italics* = direct quotes from the text of the test; '. . .' = speech (possibly abbreviated) by participant; /1. = one of the alternatives in the multiple-choice question; [...] = insertion in speech.

10. We do not advocate this as a teaching method but only as something to be used for difficult passages.

11. Evans (1988) also discusses problems in EFL reading in connection with students' academic careers. He ascribes problems to deficient understanding of syntax.

12. Generally speaking there is a lack of empirical studies of reading which could be used. It is also doubtful if results from mother tongue reading can be applied uncritically to foreign language reading (cf. also Block, 1992) - in this context the vocabulary and syntax problems alone seem to be formidable barriers.

The lighthouse of Anholt, Denmark

APPENDIX 1

Appendix 1: The first page of Sprogtest

År	Md	Inst.	Fag	Hold	Prsnr.	Test
						2 3 0
1	3	5	7	11	15	18

Sæt x ved det rigtige svar. s.1
HUSK AT BESVARE ALLE SPØRGSMÅL. [s.1]

Volunteer miners kept coal-mines safe

(21)
1. carried out
Voluntary safety work 2. obstructed ☐ by miners during their strike in 1974 meant
3. neglected

that all coal mines could resume production when the strike ended. After the previous strike 25 out of about 800 coal-mines were lost, but this time co-operation between miners and management ensured that the pits were

(22)
1. operating
2. kept safe ☐ during the strike. In addition to deputies and inspectors, who
3. closed

reported for duty each day, many miners

(23)
1. were forced
2. refused ☐ to help in taking safety measures with their union's approval. A
3. volunteered

Coal Board spokesman said: 'The vast majority of coal mines was standing up well. Precautions were taken

(24)
1. in spite of
2. in disregard of ☐ experience gained during the previous
3. in the light of

(25)
1. experts and administration
strike. Each pit has its own 2. peculiarities and weaknesses ☐ and it is the men who
3. output and production

know them, at local level, who went down

(26)
1. where necessary
to see what was happening, and 2. not ☐ to take action
3. to discuss whether

(27)
1. administrative problem
The main 2. trouble for them ☐ was that in some pits the
3. reason for mining

Appendix 2 and 3

Appendix 2: Distributions of answers in multiple-choice questions in Sprogtest – Text 230 (see Appendix 1)

	Frequency	%	Valid %
Question 1			
Alternative 1	1274	78.8	79.9
Alternative 2	183	11.3	11.5
Alternative 3	138	8.5	8.7
Missing cases	21	1.3	Missing
Total	1616	100.0	100.0
Valid cases: 1595			
Missing cases: 21			
Question 2			
Alternative 1	181	11.5	11.6
Alternative 2	1093	67.6	68.4
Alternative 3	320	19.8	20.0
Missing cases	17	1.1	Missing
Total	1616	100.0	100.0
Valid cases: 1599			
Missing cases: 17			
Question 3			
Alternative 1	220	13.6	13.7
Alternative 2	278	17.2	17.4
Alternative 3	1102	68.2	68.9
Missing cases	16	1.0	Missing
Total	1616	100.0	100.0
Valid cases: 1600			
Missing cases: 16			

Appendix 3: Feedback in Sprogtest (sample)
Year: 91 *Month*: 10 *Institution*: 01 *Discipline*: 4331 *Group*: 0108

Person number	Text	Score*	Value (weighted score)**	Assessment
7	230	13/13	21/21	Excellent
	330	17/17	20/20	Excellent
	420	14/15	17/19	Excellent
15	230	11/13	20/21	Excellent
	330	14/17	17/20	Good
	420	12/15	15/19	Excellent
17	230	12/13	21/21	Excellent
	330	15/17	18/20	Excellent
	420	10/15	13/19	Good
		You have trouble understanding syntax		
22	230	9/13	15/21	Good
	330	15/17	20/20	Excellent
	420	11/15	9/19	Good
		You have trouble understanding syntax		

*Number of correct answers/maximum correct answers.
**Number of points scored/maximum points possible.
Note: Missing 'person numbers' merely imply that the intervening tests were not used, for instance because they happened to be surplus tests.