DOCUMENT RESUME

ED 281 387 FL 016 664

AUTHOR Oscarson, Mats
TITLE Native and Non-Native Performance on a National Test in English for Swedish Students. A Validation Study. Skrifter fran Avdelningen for sprakpedagogik 4. Report No. 1986-03.
INSTITUTION Gothenburg Univ. (Sweden). Dept. of Educational Research.
PUB DATE 86
NOTE 138p.; Appendices contain pages of small, light print.
AVAILABLE FROM Department of Education and Educational Research, Gothenburg University, Box 1010, S-431 26, Molndal, Sweden.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS Comparative Analysis; *English (Second Language); English for Academic Purposes; Foreign Countries; Grammar; *Language Proficiency; *Language Tests; Listening Comprehension; Native Speakers; Reading Comprehension; Secondary Education; *Standardized Tests; Syntax; *Test Validity; Vocabulary; *Writing Skills
IDENTIFIERS *Sweden

ABSTRACT

A replication of a previous study assessed the construct validity of a national test for academically oriented Swedish upper secondary students. The analysis consisted of a comparison of test results of 10 percent of the Swedish test population to results obtained from a sample of native English speakers of the same age. Analyses of the Swedish students' English written production skills were also performed. The test battery included subtests of vocabulary, phrases, grammar, reading comprehension, and listening comprehension. Results for the native English-speakers were significantly higher on all subtests but one, a reading test, suggesting that the test is a valid measure of English language proficiency. Another important observation was that the native and non-native score levels were unevenly distributed on the various subtests, interpreted as a sign of variable subtest validity and a need for modification of test content and format. Analysis of the students' English writing skills revealed little correlation between native and non-native average scores, interpreted as a difference in language ability in the two groups. Certain tasks appeared to be much more difficult for the non-native students. (MSE)

NATIVE AND NON-NATIVE PERFORMANCE ON A NATIONAL TEST IN ENGLISH FOR SWEDISH STUDENTS

A Validation Study

Mats Oscarson

Skrifter från Avdelningen för språkpedagogik 4

Report No. 1986:03
Department of Education and
Educational Research
Gothenburg University

Department of Education and Educational Research
Language Teaching Research Unit
Gothenburg University

NATIVE AND NON-NATIVE PERFORMANCE ON A NATIONAL TEST IN ENGLISH FOR SWEDISH STUDENTS

A Validation Study

Mats Oscarson

Skrifter från Avdelningen för språkpedagogik 4

Report No. 1986:03

ABSTRACT

Oscarson, Mats
Native and Non-Native Performance on a National Test in English for Swedish Students: A Validation Study

Report No. 1986:03

ISSN 0282-2156
Number of pages: 147

The present report describes a replication of a previous study (Oscarson, 1986) which aimed to assess the construct validity of the 1983 version of the National Test ('Centrala provet') in English for the academically oriented part of the Upper Secondary School. The chief aim of the present study was to investigate the construct validity of the 1985 version of the same test. The method employed was a quantitative (i.e. statistical) analysis of the results obtained by a group of native students who had been asked to take the test. The assumption behind the experiment was that educated native speakers ought to be able to reach very high scores on an English proficiency test which has claims to high validity. Another point of departure was the belief that native speakers ought to be able to outperform non-natives in all the different areas of language which a valid test sets out to measure.

A subsidiary aim was to study Swedish students' results on open-ended ("free-form") items in the test in order to evaluate some aspects of the written production skills generally attained in the Upper Secondary School. This qualitative investigation, which involved detailed linguistic analyses of answers, was of special interest because of the fact that the sample could be regarded as representative of the whole student population.

The native group consisted of 166 English students at a Sixth Form College in Manchester. They were all of the same age as the Swedish students and represented a cross-section of the student population in respect of academic and linguistic ability.

The Swedish group consisted of a random 10% sample of the total population of 34,000 students tested in 1985. The experimental sample thus comprised 3,400 students. In the analysis of the open-ended items, a random sub-sample of 176 students was used.

The test battery included sub-tests measuring vocabulary, phrases, grammar, reading comprehension, and listening comprehension. The total number of tasks was 100.

It was found that the native students obtained significantly higher results on all sub-tests but one (a Reading Test). Their average rate of correct scores was 83%. The corresponding Swedish score was 61%. The relatively high native performance level was taken to warrant the conclusion that the National Test is a valid measure of English language proficiency. The outcome of a separate analysis of the scores achieved by an academically more advanced group of native students provided further support for this conclusion.

Another important observation was that the various sub-tests resulted in very unequal relationships between native and non-native score levels. The most pronounced difference appeared on the Vocabulary test, whereas a sub-test measuring extended reading (passage comprehension) produced no difference at all between English and Swedish students. Only a relatively small difference was found on the Listening test. While some variation of the size of the relative difference between sub-test scores was expected (due to natural differences in skills profiles), these results were interpreted as signs of variable test validity, and this led to certain suggestions for modification of test content and format.

With regard to the secondary objective of the study (the qualitative analysis) it was found that there was hardly any correlation between native and non-native average scores on the productive items. This was understood to signify a difference in the structure of language capacity between the two samples. Certain tasks, notably those involving the use of idiomatic phrases, were markedly difficult for Swedish students. The same was true of certain points of grammar, whereas overall comprehension appeared to be quite well developed.

The report contains concrete and detailed linguistic analyses and exemplification of answers delivered by native as well as non-native students and may therefore usefully serve as a resource in teacher training and similar contexts.

The report can be ordered from:

Department of Education and Educational Research
Gothenburg University
Box 1010
S-431 26 Mölndal
SWEDEN

ACKNOWLEDGEMENTS

Research reports can usually be seen as manifestations of sustained cooperative efforts and this one is no exception. In the course of the planning and implementation of the investigation, as well as in the preparation of the report itself, I have benefited from the assistance and advice of many people to whom I am indebted. In particular I should like to express my gratitude to the staff (Peter Birch, Ann Dowling, M.Galvin, G. Griffin, W. Moran, K. O'Kelly, Ann Williamson) and students at Xaverian Sixth Form College in Manchester for the willingness and determination with which they participated in the assessments. I am also very grateful to Mike Kelly of the Centre for Educational Development and Training at Manchester Polytechnic who undertook the demanding task of coordinating the testing sessions and also provided me with valuable background information about the groups involved.

Furthermore I am indebted to the following colleagues at Gothenburg University for reading and commenting on some or all parts of the manuscript during various stages of development: Ingvar Carlsson, Nils-Henrik af Ekenstam, Göran Eriksson, Rigmor Eriksson, Jan Hellekant, Eric Kinrade, Torsten Lindblad, Pat Nilsson, Torborg Norman, Sölve Ohlander, and David Wright. Their contributions have been much appreciated.

Finally my thanks are due to Inga-Britt Holmgren for the final editing of the typescript.

CONTENTS

TABLES

FIGURES

1 INTRODUCTION

The work described in the present report forms part of a long-term research and development programme which has recently been linked up with the administration of the national standardized tests in the Secondary and Upper Secondary schools in Sweden. The programme was initiated with a view to ensuring a scientifically sound basis for the National testing methods currently used. Another important aim was to see to it that better use was made of the large amounts of statistical data which are amassed each year as a result of the assessments. The data may be used, for instance, for purposes of evaluating the effects of teaching investments or for purposes of monitoring the results of changes in educational policy.

The nationwide tests (in Swedish, 'standardprov' and 'centrala prov') cover several subjects and are administered at various points in the Secondary and Upper Secondary school (from grade 8 onwards). Foreign language tests are, at present, administered in grade 8 (English tests) and in the second year ("grade 11") of the Upper Secondary School (English, German, and French tests). Other subjects tested are Mathematics, Physics, Chemistry, and Swedish.

The chief object of the measurements is to make it possible for teachers to compare the proficiency levels of their classes with the average national levels. Being able to do this is important because of the grading system used in Swedish schools. Grades are awarded on a 5-point scale and are distributed on a statistical basis (in the Upper Secondary School in the proportions of 7-24-38-24-7 per cent of the population for grades 1, 2, 3, 4, and 5 respectively, grade 5 being the highest). Accordingly, the results on the national tests are interpreted in norm-referenced terms, i.e. in relation to the performance of the entire student population taking the same test (and following the same course of study).

Grading in the individual class is adjusted so as to conform to the general outcome of the National Test (but it is still the teacher who makes the final decisions on the distribution of the various grades). The main aim of the whole testing operation is of course to ensure that a given grade can be taken to mean approximately the same thing wherever it is awarded, or, seen from a slightly different angle, to ensure that students receive fair treatment in terms of assessment and grading, regardless of what school they happen to attend. It should be added that there are no final examinations in Swedish schools.

All standardized National tests are extensively pre-tested and subjected to careful scrutiny by groups of experts (including teachers, test constructors, and administrators) before they are moulded into their final form. Shortly after the day of the test, the results of a few thousand students are collected by random sampling, and norms are calculated and fed back to the schools to be used as guidelines when students are being graded at the end of the term.

More detailed information about the principles of assessment and evaluation in Swedish schools is given in the official document "Assessment in Swedish Schools", which may be obtained free of charge from The National Swedish Board of Education, The Information Section, S-106 42 Stockholm. A description of language testing in Sweden as seen from an outside observer's point of view is given in Orpet (1985).

An attempt at validating the current tests in English was also made in 1983. The procedures and results are summarized in Section 3.1 of this report. The present study is a replication of that earlier study. Similar work has been undertaken in two more languages, French and German. The results have been described by Jan Hellekant (for French) and Nils-Henrik af Ekenstam (for German) in separate reports from the Language Teaching Research Unit, Gothenburg University. A brief résumé of the main findings is given in Chapter 3.

2 VALIDATION OF LANGUAGE TESTS: SOME BASIC CONSIDERATIONS

Test validation may be broadly defined as the process whereby the outcomes of a test are assessed in relation to the purpose of the testing. Applying this definition to language testing, we may then say that validation is a matter of determining to what extent a given test yields information about the testees' capacity for functioning in the language according to certain predetermined linguistic criteria. The criteria may take the form of a set of language learning objectives laid down in a syllabus (as in a school situation), a job requirement specification, a stipulated level on a descriptive language proficiency scale, etc. A test which truly samples a body of criteria of this kind (i.e. criteria about which conclusions are to be drawn) is said to have content validity.

The validity of a language test may also be assessed in relation to a theory of what it really means to know a language, e.g. in the form of a specification of the various abilities and traits which together constitute the more general psychological concept (or "construct") of language proficiency (see for instance Bachman and Palmer, 1982, de Jong, 1983). A specification of this kind may involve the use of descriptive categories such as 'mastery of the phonemic system', 'word recognition', 'verbal reasoning', 'retention of information', 'strategic competence', 'grammatical competence', 'sociolinguistic competence' etc. The construct of, for instance, understanding spoken English, may perhaps be thought of in terms of statements such as the following: 'The proficient person has control of the phonemic system of the English language and is able to identify and interpret all important stress and intonation patterns ... He can make relevant distinctions between morphological and lexical units ... His tolerance to reduced redundancy caused by interference in the

channel is such that ... When confronted with a sample of spoken English he is able to extract from it the same information as other listeners of a comparable experiential and educational background ... It may be predicted that he will obtain high scores on other accepted measures of listening comprehension ...' etc. (It should be emphasized that these statements have been formulated only for the sake of exemplification of a principle; they do not constitute a definite proposition.)

The next step in a construct validation procedure is to investigate to what extent the test under consideration measures the construct, or constructs, hypothesized - for instance by studying jointly the intercorrelations of this test and a number of others. If the test yields scores that accurately describe testees in terms of the relevant constructs, it is said to have construct validity. (For in-depth treatment of the principles of construct validation, see for instance Thorndike and Hagen, 1969, and Cronbach, 1971.)

It might be added, in passing, that there has been some disagreement as to the extent to which linguistic competence is divisible into separate components. Some experts on testing, notably Oller (1979), have argued in favour of a unitary competence model which postulates a common one-dimensional trait (a general component or factor) that explains all of, or most of, the variance in any language test. This so-called indivisibility hypothesis can now be said to have been disproved by other researchers (cf for instance Sang et al, 1986) and Oller has since modified his position.

Several other varieties of test validity have been identified, e.g. face validity, which refers to the extent to which a test appears to be a valid measuring instrument (especially in the layman's view), concurrent validity, which relates to the question of whether tests that supposedly measure the same skills actually correlate statistically with each other, and predictive validity, which refers to the accuracy with which a

test predicts future job or educational performance. Both of the latter types of validity are arrived at by comparing the test results with some independent criterion measures, and are often subsumed under the more general term criterion-related (or empirical) validity.

The type of validity with which we are concerned in the present series of investigations may be classified as construct validity, although we do not start from a hypothesis of what particular concepts or constructs our tests are supposed to measure. Instead we work on the assumption that the conglomerate of abilities that make up what we ordinarily call general language proficiency must be possessed, to a very high degree, by native speakers of the language and that non-native speakers cannot be expected to possess the same degree of ability as native speakers do. Many other researchers have endorsed this approach to test validation. Oller (1979), for instance, holds the view that "... native performance is a more valid criterion against which to judge the effectiveness of test items than non-native performance is" (p 203). He goes on to say that

> "In a fundamental and indisputable sense, native speaker performance is the criterion against which all language tests must be validated ... The choice of native speaker performance as the criterion against which to judge the validity of language proficiency tests, and as a basis for refining and developing them, guarantees greater facility in the interpretation of test scores, and more meaningful test sensitivities (i.e. variance)" (p. 204).

A further assumption underlying the present study is that non-native speakers, i.e. learners, will not have advanced equally far in the various domains of language proficiency. "Artificial" learning in a formal educational context is likely to favour the development of certain abilities more than others and hence one can expect differences between native and non-native speakers to vary in accordance with the types of task involved. It may be predicted, for example, that sub-tests measuring non-specialized reading comprehension skills will

result in relatively high non-native scores (text-based materials and exercises being very prominent features of foreign language instruction). Likewise, one may predict that there will be a sharper contrast in native and non-native performance on tests measuring comprehension of everyday spoken English than on tests measuring comprehension of, for instance, formal speech.

Lastly we assume that it is possible to control factors other than linguistic (situational, motivational etc) which may have an influence on performance when the test is administered under different conditions and in different settings (in our case classes and classrooms in Sweden vs. other countries).

The following chapter describes earlier experiments geared to the type of construct validation discussed above.

3 PREVIOUS EXPERIMENTS

3.1 Validation of an English test

The experiment to be described in the present report is a replication of an earlier validation study which will be briefly reviewed here. For a more detailed account of procedures and results, see Oscarson, 1986 (published in Swedish with a summary in English).

The main aim of the experiment was to investigate, by means of an analysis of native English students' performance, the validity of the 1983 version of the National Test in English for the Upper Secondary School in Sweden. Another aim was to exemplify, in concrete terms, the level of proficiency in English of a representative sample of the target group, i.e. students in the second year of the "theoretical" three- and four-year lines of the Swedish Upper Secondary School. The native English group comprised 105 A-level students at three Upper Secondary schools in York, England (two comprehensive schools and a grammar school). The Swedish group consisted of a random (i.e. representative) sample of 3,300 students drawn from the entire population of approximately 33,000 students who took the test in 1983. The average age of the students in both groups was 17.

The validation study was based on the premise that the level of mastery of the language was considerably higher in the native group than in the non-native group and that a valid language test would disclose this real difference in ability very clearly. Accordingly, a very small difference in test results would be taken to indicate inadequate test validity, at least in one sense of the term (i.e. that of construct validity; cf Chapter 2).

The test was, as far as possible, administered under comparable conditions in the two groups. The same instructions (written in English) were used, the time allowed for the various sub-tests was the same, and the same criteria for marking were applied. The English students were less well acquainted with the testing techniques than the Swedish students, naturally enough, but this difference did not affect the general outcome of the comparison as far as could be ascertained. Motivation was high in both groups.

The testing sessions in England were organized and supervised by an English teacher trainer who is fluent in Swedish and familiar with the Swedish educational system (having spent some time at a Swedish university).

The main results are summarized in Table 1 (for a description of test content, see Section 5.0 in this report):

Table 1 The Results of the 1983 Validation Study (York): Mean Scores in Native and Non-Native Groups (N = 105 and 3,300 respectively)

	Sub-test and Subjects	No. of Items	Mean Score ($\bar{x}$)	Standard Deviation (s)	Mean Score in % of Max Score	Reliability (KR 20)
1:1	Voc-Gr A	15				
	Native students		14.57	0.64	97.1	.03
	Non-Native st.		8.51	3.15	56.7	.67
1:2	Integrative T.	37				
	Native		32.20	3.91	87.0	.78
	Non-Native		22.24	7.97	60.1	.88
2:1	Reading Compr.1	11				
	Native		8.96	1.65	81.5	.48
	Non-Native		6.72	2.62	61.1	.68
2:2	Reading Compr.2	9				
	Native		7.80	1.71	87.6	.76
	Non-Native		4.98	2.12	55.3	.57
2:3	Reading Compr.3	9				
	Native		7.85	2.07	87.2	.86
	Non-Native		5.22	2.41	58.0	.70
3:1	Listening Compr.	10				
	Native		7.41	1.66	74.1	.46
	Non-Native		6.43	1.97	64.3	.45
3:2	Voc-Gr B	14				
	Native		11.91	1.87	85.1	.55
	Non-Native		9.39	3.32	67.1	.78
	TOTAL	105				
	Native		90.79	7.77	86.5	.84
	Non-Native		63.49	19.40	60.5	.94

The validity of the test was thus high according to the criterion (successful native performance). The average native correct response rate was 87 per cent of the maximal total score. The corresponding Swedish figure was 61 per cent. The pattern of scores is illustrated in Figure 1:

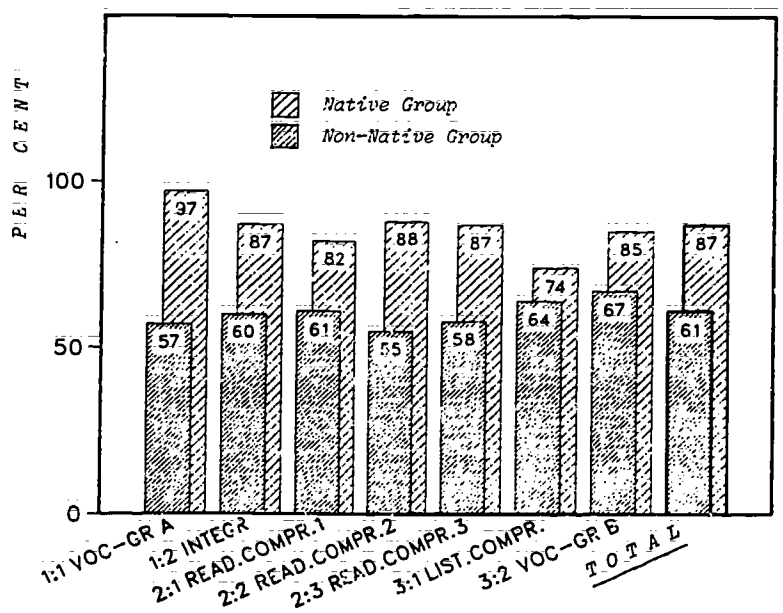


Figure 1 The Results of the 1983 Validation Study: Proportion of Correct Responses per Sub-Test

As regards the levels of mean scores, it should be pointed out, firstly, that the test in question is quite advanced as it is used in order to gauge the proficiency of students in their eighth or ninth year of instruction in the language. Secondly, the test is of the norm-referenced type (cf Chapter 1) and it is, for this reason too, pitched at a relatively high level of ability in order to yield an optimal spread of individual results. The fact that the native score was less than perfect is partly explained by these circumstances. The parallel experiments with French and German tests (see below), which are less advanced, resulted in much higher native scores, whereas the non-native scores were in the region of

55-65 per cent of the maximal score, much as in the English investigation. It might also be added that this particular level of average correct scores is intentional and has to do with the fact that test data are used for norm-referenced, rather than criterion-referenced, interpretation. (See further Section 8.1.1).

Another notable finding was that the various sub-tests resulted in very unequal differences between the two groups. This was taken to indicate that the degree of validity varied with test type (although certain discrepancies between native and non-native mean score patterns were expected). The most clear-cut differences were obtained on sub-tests that essentially measured command of lexis and understanding of idiomatic sentences (1:1 and 2:2). The smallest difference was obtained on the sub-test measuring listening comprehension. The distributions of native and non-native scores overlapped to a very large extent (one third of the Swedish students outperformed half the native students). The result made a replication with a different version of the test highly desirable.

The study further seemed to confirm other research results (e.g. Löfgren, 1969; Carroll, 1973; Angelis, 1977) which have pointed to a strong correlation between certain psychological factors, such as deductive ability and short-term memory, and ability to answer multiple-choice questions on the contents of texts (written or spoken). Such tasks are set in sub-tests 2:1 (Reading Comprehension) and 3:1 (Listening Comprehension).

The analysis of results on individual open-ended items suggested that Swedish students' <u>formal</u> command of English is comparatively weak. Elementary mistakes in lexis and syntax were not uncommon (whereas they very rarely appeared in the native group). In contrast, the <u>functional</u> command of the language (again seen in relation to the performance of the native speakers) was quite good. It was assumed that the observations needed further substantiation.

Finally it may be noted that it was possible to use the test under investigation for criterion-referenced interpretation, to some extent, in spite of the fact that it was primarily designed for purposes of norm-referenced evaluation. The reason was that it contained productive parts, in addition to the multiple-choice parts (the two types of task being represented in about equal measures).

3.2 Validation of a French test

Two parallel experiments investigating the validity of a French and a German test were undertaken in 1985. In the French experiment, the 1985 version of the National test in French was administered to 120 pupils aged 14-16 at a non-selective secondary school (a collège) just outside Lille, France (for details of this experiment, see Hellekant, 1986, in Swedish with a summary in French). Their results were compared with the results obtained by a random sample of 200 Swedish students belonging to the group for whom the test was constructed (i.e. 17-year-olds in the second year, "grade 11", of the Upper Secondary School). The reason why a younger student sample was used in this experiment than in the English study described above was that the French test is a great deal easier than the English test (French being the students' second foreign language, as well as being considerably more difficult than English for speakers of Swedish); 17-year-old native speakers of French would probably have found the tasks boringly simple (which might have jeopardized the validity of the research).

Testing procedures and materials were the same in France as in Sweden. The French students were of course given instructions in their own language (as were the Swedish students).

The results may be summarized as follows (following the author):

On average, the French students mastered 92.6 per cent of the test items which was interpreted as a very satisfactory sign of test validity. The corresponding Swedish percentage was 55.7. The native students reached their highest average score (98 per cent of the maximal number of points) on a sub-test which consisted of a dictation. It should be added that the sub-test was designed in such a way that it measured, first and foremost, accuracy of spelling and little else.

The second highest native score (94.5%) was recorded in the listening comprehension part of the test; the outcome was interpreted as refutation of a certain amount of criticism that has been levelled at this part of the test (concerning rate of speech, dependance on memory etc).

Extremely high native scores (99%) were attained in two sub-sections of a sub-test measuring knowledge of grammar, words, and phrases by means of multiple-choice tasks. Other sub-sections measuring the same domain by means of written production tasks (gaps to be filled in) resulted in much lower average scores (ranging from 83.5 to 89% of the maximal score). The author's conclusion is that there is a need for revision of the marking criteria and that a wider tolerance to certain deviations from the traditional linguistic norm must be shown if we want to assess practical communicative skills in a reliable way.

The mean performance level of the native students on the sub-test measuring reading comprehension was relatively low (89%). The author of the report expresses some concern that too strict demands may sometimes be made on students' ability to draw logical conclusions on the basis of facts presented in pieces of text of some length. Cutting up long texts into shorter segments, each followed by a set of questions, is recommended.

3.3 Validation of a German test

In the German experiment, the 1985 version of the National Test in German was administered to 500 students at seven Upper Secondary Schools in the Federal Republic (for a detailed report on this experiment, see af Ekenstam, 1986, in Swedish with a summary in German). Six of the schools were so-called Gymnasien, which means that their students can be described as a very select group as far as academic ability is concerned; the vast majority of the students at these schools are preparing for higher education at university level and they represent only a quarter of the entire age group. The seventh school was a Gesamtschule, which is a type of school attended by students of much more varying academic ability. (The test results did not, however, come out differently, generally speaking, in the two types of school.)

As in the experiments described above, the testing procedures and materials were equivalent to those used with Swedish students. Instructions were translated into German. In Sweden, the test is taken by students in their fifth year of instruction. The German language is considered relatively easy (being fairly closely related to Swedish), and is the most commonly chosen second foreign language in Swedish schools. The general proficiency level reached is normally higher than that in French (which is more difficult for speakers of Swedish) but lower than that in English.

On average, the German students solved as many as 97% of the test items correctly according to the key. The author concludes that this very high native score is an indication of very satisfactory construct validity in that it shows that superior proficiency in the language leads to successful test performance. Skills that are less developed result in significantly lower test results, as evidenced by the average non-native score, which was 84.3% of the maximal number of points.

The German students reached their best result on a sub-test measuring Grammar, vocabulary, and phrases; the success rate was as high as 98.2% of the maximal score. The second best result was obtained in the Listening Comprehension part (97%), while the Reading Comprehension score was comparatively low (93.2%).

In the report, the author also discusses individual items, viz. those which resulted in scores which were substantially lower than average in the native group, and he draws some conclusions for future test construction on the basis of the experience gained.

3.4 Summary and conclusions

The experiments reported above were very illuminating in several respects. To begin with they all showed that testees who have native competence normally obtain very high test scores; hence the tests may confidently be regarded as valid in the sense stated in the objectives. Particularly sub-tests measuring control of the elements of language (words, phrases, grammar) by means of the multiple-choice technique proved to be highly sensitive to the sort of indisputable linguistic capacity that native speakers possess.

Other findings were somewhat less reassuring. Thus the validity of certain sub-tests seems to be open to some doubt. The most questionable case was the Listening Comprehension part of the English test in which the native speakers no doubt experienced problems at times. To a lesser extent this was also true of the Reading Comprehension parts of both the English and the French tests.

The qualitative examination of the English test results seemed to lay bare a weak spot in the Swedish students' command of the language, viz. in the area of formal accuracy. Further research into this problem is needed.

These facts taken together called for a renewal of the experiment in England (using a different version of the test). The work was undertaken in 1985 and will be described on the following pages of this report.

4 OBJECTIVES

The most important aim of the replication study was to obtain further empirical evidence as to the validity of the National Test in English currently used in the Integrated Upper Secondary School in Sweden. A subsidiary aim was to investigate some aspects of language performance of a representative sample of Swedish students taking the test and to compare the performance of Swedish students with that of a similar native sample.

The main aim was to be achieved by means of a quantitative (statistical) analysis of the performance of native English students on the test in comparison with the results obtained by the non-native speakers. The assumption underlying the experiment was that for a test to be valid it ought to be possible for native speakers to reach significantly higher scores than non-native speakers of a comparable social and educational background (although probably to varying degrees in the various skills). In other words, it was posited that, all other things being equal, the average native speaker is palpably more competent and proficient than the average non-native speaker and that this applies to all areas of language use which a valid test is designed to measure. Expressed in more technical terms it was assumed that native speakers possess the construct (cf Chapter 2) of English language proficiency to a much higher degree than comparable non-native speakers. If a test which purports to measure English language proficiency does not register this difference reasonably clearly one can suspect that there is a certain lack in test validity (i.e. construct validity); the test may still have face validity (i.e. it may "look good") or have content validity (i.e. it may "test what has been taught").

Moreover, as has often been noted in the literature, any valid foreign-language test ought to be a test in which the educated native speaker can obtain an almost perfect score (i.e. a score approaching 100% correct response rate):

> " ... if the test is administered to native speakers of the language they should make very high scores on it or we will suspect that factors other than the basic ones of language have been introduced into the items" (Lado, 1961:323).
>
> "Any foreign language test should be a test in which the educated native speaker can obtain a perfect score" (Klein-Braley, 1985:83).
>
> "Natives should always be criterial in a test item, i.e. they should (90% of the time at least) get it right" (Davies, 1985:103).

The validity and reliability considerations were therefore extended to the item level.

The subsidiary aim was to examine certain productive language skills acquired by a representative sample of students in the second year of the "theoretical" (i.e. academically oriented) Upper Secondary school in Sweden. The skills area was limited to the production of lexically and syntactically acceptable written English, as documented by the outcome of sub-tests 1:2 and 3:2 (cf Sections 9.2 and 9.3). Comparison was to be made with language samples produced by the native speakers. Basically, then, the purpose was to describe ability levels in absolute terms by means of a detailed qualitative analysis of concrete answers to individual open-ended test items. The work was possible to undertake because of the fact that the test includes "open-ended" tasks in addition to tasks of the multiple-choice format. The test is, however, primarily designed for norm-referenced interpretation purposes (cf Introduction).

It should be emphasized that the qualitative analysis of the responses in the Swedish group was of particular interest in view of the fact that the subjects constituted a random sample of the entire population of students for whom the test is designed, i.e. the sample represented, in every important

respect, all students (or nearly all since there was a small percentage who did not take the test) in the second year of the Swedish non-vocational Upper Secondary School. This means that the investigation made it possible to survey typical error patterns and their frequency in the population and to identify weak and strong points in the students' command of the language. It is, thus, by virtue of the representativity of the material that the present analysis of errors merits some special attention. The errors as such may not be very interesting; they are probably all too familiar to any teacher of English, at least in the Scandinavian context.

5 DESCRIPTION OF THE TEST

5.1 Function

The test that was used in the study was the 1985 version of the National Test in English for second-year students in the academically oriented non-vocational Upper Secondary School in Sweden. The format of the test has remained unchanged for a number of years, but the contents (stimulus material and tasks) are completely renewed every year. As was explained in the Introduction, the prime function of the test is "calibration", i.e. the test results are used as an aid by means of which teachers may, or indeed should, adjust their standards of grading to what turns out to be average national performance levels for the various grade categories. The test is administered on the same day in all schools throughout the country.

The test is thus of the norm-referenced (rather than the criterion-referenced) type. Nonetheless certain parts of the test (cf Chapter 4) lend themselves to criterion-referenced interpretation, i.e. the results can be used in order to describe the testees' language skills in absolute as well as in relative terms.

5.2 Contents

The 1985 version of the test consisted of the following parts (with sample items):

1:1 Vocabulary Test (18 items, multiple choice)

Examples:

(1)

PEG:	You can't get in if you're under eighteen.	A go the way B put it up
BOB:	I'll stick on a moustache and tell them I'm twenty.	C clear it D come off
PEG:	You won't - - - - . Not with that baby face of yours.	E get away with it

(5)

The disputes were often heated and on one - - - I remember the meeting broke up in disorder.	A. event B. incident C. occasion D. occurrence E. opportunity

Most of the items (12 in all) tested single verbs or verb phrases, three tested adjectives and three tested nouns.

1:2 Integrative Test (35 items, a running text with one-word gaps to be filled in; further details about test content are given in Section 9.2)

Example:

KEN: Tell me a little about your family, Pam. For instance, what _____(1)_____ your dad do?

PAM: He's an engineer. His job takes him all over the country, and abroad, too, sometimes, so he's hardly ____(2)____ at home. Mother says it's like ___(3)___ married to a sailor.

KEN: Yes, I can imagine ...

...

The Integrative Test, which exemplifies the so-called cloze procedure (Taylor, 1953; Oller, 1979), is reproduced in full in Appendix 2.

2 Reading Comprehension Test

Part 1 (12 items, multiple choice, comprehension questions on a text comprising approximately 1,200 words)

Example (the first paragraph of the text):

This is a newspaper article written by a British journalist called Joan Wilson.

The trouble with abroad is that you are liable to come up against unpredictable obstacles. In Paris I once wanted to find the dialling code for England, which I thought would be done in the twinkling of an eye. But it took me ages. I tried looking up 'Angleterre', then looked under 'Grande Bretagne' and drew another blank. Only after considerable brain cudgelling and much irritation did I hit upon 'Le Royaume Uni'. And if tracing the name of your own country can be hard, trying to work out what any country calls its own railway system is next to impossible. Either you know it or you don't, and if you don't there's no ringing the station to find out the time of the train.
...

1 Ms Wilson points out that in a foreign country ...
  A. telephone directories are often misleading
  B. you may easily run into unexpected difficulties
  C. there is usually no information service at railway stations
  D. the railway system is mostly very complicated

The questions were interspersed in the text in groups of three or four at a time, that is, the text was broken down into smaller sections, each followed by a set of questions.

Part 2 (10 items, multiple choice, each consisting of a "mini-text" with a one-word gap)

Examples:

(13)
After seeing some extremely violent porno-horror movies, I decided that I would welcome some sort of legislation which would - - - the general distribution of video "nasties".

A. prevent
B. produce
C. develop
D. enlarge
E. lighten

(18)
They are trying to make Mr Dawson renounce his position. However, his associates are emphatic that he will not go - - - . They say he has dug in his heels and is not the resigning type.

A. against them
B. mad
C. back
D. quietly
E. under protest

3:1 Listening Comprehension Test (11 items, multiple-choice, comprehension questions based on an audio-taped dialogue)

Example:

(Tape) This scene takes place in a coal-mining village. At the nearby pit the miners are on strike. Meg is in her own home, and she is talking to an older man, Thomas, who has just come in.

MEG: Well, what happened?
THOMAS: We're staying out.
MEG: What was the voting?
THOMAS: Show of hands. It was obvious.
MEG: So nobody counted them.
THOMAS: They don't count at pithead meetings. You know that. Not unless it's close.
MEG: Yes, I know that. And I know the shop-stewards see what they want to see.
THOMAS: Have you got a cup of tea, Meg?
MEG: I might have. Where's Dai?
THOMAS: He's with some of the boys. He'll be here in a minute.
MEG: I suppose he voted for the strike, too. I expect you lectured him all the way to the pit.
THOMAS: Meg, will you just tell me ...
MEG: Will you just tell me how we're going to manage over Christmas? And how we're going to pay the mortgage. The mortgage on our house, mind you - Dai's and mine.
THOMAS: You asked me to live here.
MEG: Yes, I did. And most of the time I'm glad I did. It's just that ...
THOMAS: What?
MEG: Thomas, I didn't ask you to bring union politics with you, that's all.
THOMAS: You want Dai to be another soft one like the rest? The bosses crook their fingers and my son comes running - is that what you want?
...

(Tape) Question No. 1:
Where are Meg and Thomas? (Repeated once)

(Test booklet)
A. In a workshop
B. In a café
C. In Meg's home
D. In Thomas's home

(Tape) Question No. 2:
What is worrying Meg? (Repeated once)

(Test booklet)

A. She fears they'll run out of money
B. Her husband is out drinking
C. The extra work Thomas gives them
D. Thomas's soft attitude about the strike

The scene was recorded in a studio in London. The parts were played by professional actors, who spoke with a slight Welsh accent in order to create a realistic atmosphere. The recording was quite lively.

3:2 Vocabulary-Grammar Test (14 items, fill-in, consisting of mini-texts each with a multiple-word gap; further details about test content are given in Section 8.3)

Examples:

(1)

JIM: This advertisement says that the machine is "foolproof". What _____ by that, Daddy?

DAD: That it's so simple that anybody can handle it, even a fool.

(4)

LEN: The damage is done
and _____
in worrying about the consequences now.

RON: That's easy for you to say.

The Vocabulary-Grammar Test is reproduced in full in Appendix 3.

The total number of items in the test was 100. About half of them (49) required active production of the students. The rest (51) were multiple choice.

The test also contained an optional written production part, an essay task, but this was not included in this validation study.

The entire test is made public as soon as it has been given in schools and is regularly reprinted in the the journal Moderna

språk, published by the the Modern Language Teachers' Association of Sweden. The version used in this experiment, including the optional essay task as well as answer keys and instructions for marking, appeared in Volume LXXIX, No. 2 (pp.174-192).

5.3 Scoring criteria

The productive parts of the test (sub-tests 1:2 and 3:2) were marked according to the following principles:

Sub-test 1:2 1 point per item was awarded for correct and acceptable words in the gaps. The point was lost if, contrary to instructions, more than one word had been inserted.

Spelling errors were penalized as follows:

1-2 errors	-
3-4	-1 point
5-6	-2 points
7 (or more)	-3 points

The minus points were subtracted from the the total score on the sub-test.

For a number of very common words (such as 'about', 'all', 'and', 'are', 'when', 'which', 'would') no variation in spelling was allowed, that is, any spelling error resulted in a 0 mark on the item in question. The total number of such words was 85. They all belong to the 100 most frequent words in the language (Svartvik et al, 1982).

The marking key contained, in addition to a list of correct answers, examples of acceptable and incorrect responses (sampled from the trial run of the test). All responses listed in the key had been checked by two native speakers (one British and one American).

Sub-test 3:2 The same as for 1:2, except that no points were taken off for spelling errors (not even spelling errors that

affected items in the list of high frequency words or spelling errors bordering on errors in grammar, e.g. 'comeing', 'geting').

The key contained specifications of possible correct answers as well as examples of acceptable and incorrect answers.

The multiple choice parts of the test (1:1, 2:1, 2:2, and 3:1) were scored on a straight 1 point per item basis. No weighting of the various sub-test aggregates was applied.

6 QUESTIONNAIRES

In addition to the test itself, two questionnaires were used in the experiment. One was directed to the native English teachers who took part in the experiment by administering the test to their students (that is, the native group). The other was directed to the Swedish teachers who, likewise, administered the test to their (i.e. Swedish) students. The English questionnaire was very brief and contained a general question on the validity of the test as well as some questions requesting background information on students and procedural matters. The Swedish questionnaire, which was quite comprehensive, included a question on each of the six sub-sections that made up the test plus a number of questions of rather more peripheral in-terest to the key issue addressed in this report. (The Swedish questionnaire was of a standard type which regularly accom-panies the test when administered in Sweden and was thus not directly devised for the purpose of our study.)

The central question asked of the participating Manchester teachers was this:

> "What is your opinion of the test itself (bearing in mind that its chief function is to assess group means)? Would you say that it is a valid measure of foreign language skills?"

The corresponding question in the Swedish questionnaire was phrased as follows (in translation):

"What do you think of the various sub-tests (testing techniques, texts, questions, individual items etc)?"

a. Vocabulary Test: ...
b. Integrative Test: ...
c. Reading Comprehension Test (the long text): ...
d. Reading Comprehension Test (mini-texts): ...
e. Listening Comprehension Test: ...
f. Vocabulary-Grammar Test: ...

Adequate space was provided for the answers. The English questionnaire is reproduced in Appendix 4.

The results will be reported in Chapter 10.

7 SUBJECTS AND PROCEDURES

7.1 The native sample

The test was administered to a total of 166 English students belonging to eight different groups at Xaverian Sixth Form College in Manchester.

The groups were so selected as to correspond, by and large, to the Swedish student population for which the test is designed, that is to say, the native students represented roughly the same type of educational and intellectual "stratum" or grouping as the Swedish students (cf Section 7.2). There is no way of knowing, however, whether the two samples can be regarded as exactly equivalent to each other in all possible respects.

A check of the results obtained by 15 students with foreign-sounding names did not reveal any large difference in ability in relation to the results obtained by students bearing typically British names. The former group, who in number amounted to less than 10 per cent of the experimental sample, scored approximately 10 per cent lower than the latter group (which in terms of overall effects may have meant a lowering of total scores by one or two per cent at most). For all practical purposes the entire experimental group of 166 students may therefore be regarded as genuinely native speakers.

The students were all 16 or 17 years old and represented different lines of study. Many of them were taking (or retaking) the 16+ examination, others were heading for O Level or A Level examinations in various subjects. The former (O level) is the ordinary school leaving examination taken at age 18, the latter (A level) is the examination required for higher education.

Seven out of the eight participating groups were described as follows by their teachers (in response to item 4 in the Questionnaire):

> How would you characterise the group/s/ in respect of academic and/or linguistic ability?
>
> "Average* group with some comprehension and spelling problems but not requiring special remedial provision."
>
> "Most of the students have GCE /General Certificate of Education/ grades 2, 3, and 4, which makes them, officially, of average ability and above."
>
> "Average"
>
> "Average ability"
>
> "Bearing in mind that the group have already failed the 16+ exam in May their academic and linguistic ability is not likely to be very high. Having gained a CSE /Certificate of Pre-Vocational Education/ 2 or 3, however, they would be slightly above the national average."
>
> "Difficult to make comparisons because although the tasks should be easy enough for these students, they are unused to being tested in this way. As students who have previously failed 16+, they are probably average to below average ability."
>
> "Generally poor. Five are on a C.P.V.E. /Certificate of Pre-Vocational Education/ foundation course, whilst the rest are retaking their 16+ English exam."
>
> (* 'Average' should be interpreted in relation to the entire population of sixth-formers in the Upper Secondary School, according to the local coordinator of the assessments.)

The eighth group consisted of 9 "upper sixth" A level students described by the local coordinator as a "bright group of above average ability . They were all 17-18 years old and were preparing for higher education at university level. As a control, the results achieved by this small group of students will be analysed separately (see Section 8.3).

7.2 The non-native sample

The Swedish group was very large (N = 3,400) and constituted a random 10% sample of the total population of some 34,000 students that took the test in 1985. The sampling was part of the yearly administrative procedures for establishing nationally valid norm data on the test. The random sampling technique plus the size of the sample guarantee that the Swedish group can be regarded as representative of the whole population of students taking the test. As only a very small percentage of students do not sit for the test (for various legitimate reasons), the large group of 34,000 students is very nearly identical with the entire population of students in the three- and four-year lines of the Upper Secondary School.

The average age of the Swedish students was 17 and they were all in their second year of the "theoretical" Upper Secondary School, which is the educational option chosen by approximately 35% of the entire age group. (About 60% choose the less academic and predominantly vocational two-year lines of study.) They were in their eighth year of English as a foreign language and had had some 500 hours of instruction (net) in the language when they took the test. Most of them were also studying German and/or French as a foreign language.

7.3 Procedures

The National Test is monolingual throughout, except for the text on the front covers of the test booklets (supplying identification data) and some back cover tabular space which teachers use when marking the test. This means that all instructions on how to take the test are in English. Consequently the only adaptation that had to be undertaken for the assessments in England was to supply an all-English front cover and to blot out two tables. When these changes had been made, the original test papers used in Sweden could be used in England as well. The adapted version of the test used in England, i.e. the one bearing an English front cover, is

exemplified in Appendix 2 and Appendix 3 (sub-tests 1:2 and 3:2 only).

In addition to the instructions contained in the test booklets, all students received oral information about testing procedures. The information given to the English students was a direct translation of the information which Swedish students received. Instructions for the English teachers who administered the test were of course also in English (cf Appendix 1).

The time allowed for the various parts was 35 minutes for each of the three sub-tests, that is 1 (including 1:1 and 1:2), 2 (including 2:1 and 2:2) and 3 (including 3:1 and 3:2). Between sub-tests 2 and 3 there was a break of 15 minutes. The total testing time (including the break) was thus 2 hours.

As far as it was possible, the tests were thus administered under the same conditions in England as in Sweden. The local coordinator in Manchester was carefully informed about the purpose of the experiment and also about the nature and function of the tests.

A total of seven native English teachers participated as administrators and invigilators. They were all provided with written information and instructions as to aims and procedures (cf Appendix 1) and also as to what information to convey to the students. The material consisted of a translation of the original instructions used by Swedish teachers.

All test materials (including test booklets, instructions, the Questionnaire and tapes for the Listening Comprehension Test) were supplied by our department. Immediately after the completion of the assessments, the materials were returned to us for marking and evaluation. The results (in the form of individual means, as well as group means per sub-section in comparison with the results obtained by the sample of Swedish students) were fed back to the staff in charge of the native groups and to the students themselves.

8 RESULTS

8.1 Preliminary remarks

By way of introduction, we will discuss very briefly a few points that may help the reader interpret the significance of the results that we are going to present. They relate to the question of the level of difficulty of the test, to the question of what types of interpretation the results allow, and to the significance of two important statistical measures, viz. the coefficients for reliability and point biserial correlation.

8.1.1 Level of difficulty

As was indicated in the Introduction, the National Test is a proficiency test (rather than an achievement test) and its general purpose is to differentiate, as clearly as possible, between students of different ability levels. In order to achieve this aim, the test must be devised to yield a maximum spread of individual results. This condition obtains when the average score is equal to half the number of tasks (=points) plus the number of points that pure guessing on the multiple-choice items would contribute. For the test under investigation the theoretical value thus calculated is 55.7; our empirical value (see below) was somewhat higher than this ideal and the distribution of scores forms a pattern which is slightly asymmetric and oriented towards the right (in technical terms, is negatively skewed). The test is, in other words, a little too easy for its purpose. It might be added, for comparison, that a test set by a class teacher in order to measure achievement during a course is normally a good deal easier.

8.1.2 Types of interpretation

Related to the above point is the question of how the test results may be interpreted. As has been pointed out several times, the National Test is first and foremost an instrument for equalizing teachers' standards of grading, and the results carry meaning primarily as comparative measures in a norm-referenced context. The items in the test have been chosen on the basis of their proven reliability and facility properties (as substantiated by prior field-testing) and not only on the basis of their suitability from a didactic point of view. (It may be added that in practice these two criteria for selection rarely come into conflict with each other.) The test may therefore very well include, for instance, a few words or idioms that the testees (or some of the testees) have not met before in their studies of English. Conversely, not all aspects of the curriculum are reflected in the structure of the test. However, these circumstances do not, generally speaking, detract from the power of the test as an instrument for norm-referenced evaluation (whereas they would if the testing were part of a criterion-referenced evaluation process).

Brief mention should also be made of the fact that the interpretation of test results must take into account certain random measurement errors which are likely to affect both individual scores and group means. Such random deviation from what might be considered the "true score" is always larger in the case of an individual student's score than in the case of a mean score calculated on the basis of the results obtained by a group of students. The reason is that in a group there tends to be some degree of balance between negative and positive random scores, which means a smaller deviation from the "true score". Computation of the so-called standard error of measurement in the Swedish group (cf Guilford 1965: 443ff), i.e. the standard error associated with the individual score, yielded a value of 4.38. This means, expressed in conventional "probabilistic" terms, that we may be 95 per cent confident

that the individual student's "true" score lies within the limits of the result obtained ±1.96x4.38 =±8.58 points, i.e. within a span of 17 points. Calculation of the standard error of the mean score achieved by a group of 25 students (cf Guilford 1965:144ff) resulted in a value of 3.58 and a confidence interval of ±1.96x3.58 = ±7.02, i.e. a span of 14 points.

8.1.3 Test reliability

Finally, a few words about the reliability indices that we will be quoting (for more details on various estimates of reliability, see for instance Guilford, 1965, or Ferguson, 1966; the following discussion is based on these sources). The reliability coefficient is basically a correlation coefficient (or, more precisely, the proportion of obtained variance of scores which is true variance) and it takes values ranging from 0 to 1. In computing the reliability of our test we used the formula known as Kuder-Richardson 20, and in a few cases a simplified form of this referred to as Kuder-Richardson 21 (estimates from the latter are generally somewhat lower than those from the former). The size of the reliability coefficient is a function of the number of items in the test and also of the size of the standard deviation, that is, the more items there are, and the greater the standard deviation, the higher the reliability coefficient is likely to be. This means that we cannot directly compare reliability coefficients calculated on tests of different lengths (they will be lower in shorter tests, all other things being equal), nor can we compare reliability coefficients calculated on tests that have resulted in very different standard deviations if, for example, this is due to the fact that either test is too difficult (i.e. has resulted in a very low average score) or too easy (i.e. has resulted in a very high average score). In both of these latter cases there is a restriction of the range of variance which has a lowering effect on the reliability coef-

ficient. It may finally be added that the reliability coefficient is in effect a measure of the homogeneity of the sample of test items (i.e. the test). This means that we will obtain the highest reliability index when the items are highly intercorrelated and measure the same trait or skill. (The two other important contributors to an optimal reliability index are equal difficulty of items and, as indicated above, maximal standard deviation).

8.1.4 Item reliability

Occasionally, reference will be made to a statistic known as the point biserial correlation (abbreviated r_{pbis}). This is a measure of the correlation between the results on an individual item and the results on all the items added together, i.e. the test score. Values may be positive or negative and vary within limits which approach -1 and +1. A high positive value indicates that those who answer the item correctly also have high total scores and, conversely, that those who fail to answer the item correctly have lower total scores. In other words, a high coefficient indicates that it is the more proficient students who master the item and this is always a desirable condition from the language tester's point of view. A low coefficient (approaching 0) tells us that the good students do no better on the item than the poor students (and this is of course unacceptable if the test is aimed to be homogeneous and valid). Finally, it should be pointed out that one must always keep the number of observations ("scores") in mind when interpreting correlation coefficients. Small numbers are usually tantamount to dubious correlations.

8.2 Summary of main results

As was expected, the native English students achieve much better results than the Swedish students. The average proportion of correct responses is 83.4 per cent in the former group as against 61.3 per cent in the latter. The difference in total test scores is 22 points.

The results are summarized in Table 2:

Table 2 Main Test Results: Native and Non-Native Groups

Sub-Test and Subjects	No. of Items and Subj.	Mean Score ($\bar{x}$)	Standard Deviation (s)	Mean Score (%)	Realiability KR20
1:1 Vocabulary	18				
Native (En)	158	16.63	1.57	92.4	.58
Non-Native (Sw)	3,409	9.67	3.72	53.7	.78*
1:2 Integrative	35				
Native	154	29.21	4.52	83.5	.82
Non-Native	3.409	21.47	7.43	61.4	.90*
2:1 Reading Compr Part 1	12				
Native	147	8.46	2.16	70.5	.57
Non-Native	3.409	8.59	3.37	71.6	.67*
2:2 Reading Compr Part 2	10				
Native	147	8.13	1.71	81.3	.60
Non-Native	3.409	5.67	2.50	56.7	.73*
3:1 Listening Compr	11				
Native	154	9.47	1.42	86.1	.47
Non-Native	3.409	7.77	1.88	70.6	.53*
3:2 Vocab.-Grammar	14				
Native	155	11.39	2.16	81.4	.65
Non-Native	3.409	8.18	3.24	58.4	.74*
Total	100				
Native	142	83.36	9.57	83.4	.86**
Non-Native	3.409	61.34	17.89	61.3	.94**

* Computed on a random sample of 172 students
** Computed according to the KR21 formula

As the table makes plain, the native speakers' superior linguistic ability shows up very clearly in the figures, and we may therefore conclude that the test is valid in the sense stated in our objectives (Chapter 4). Testees whose command of English is at an advanced level do obtain high test scores.

The reliability indices are, by and large, very satisfactory. The fact that they are generally lower in the native sample is explained by the high means, which result in a restriction of the range of variation in individual scores. In other words, the test is not difficult enough to differentiate among the best (native) students and this has the effect that the top students do not achieve higher scores than the next-to-top students, as it were. The significantly higher reliabilities in the Swedish group indicate that this problem is largely non-existent when the test is used on the home ground. Further confirmation for this conclusion is provided by Figure 6 (see Appendix 5), which shows that the individual results for Swedish students are well spread over the entire test score range, while at the same time nobody reaches the maximal score of 100 points.

Although the native students obtain significantly higher overall results, the table also shows that the size of the relative difference between the native and the non-native scores varies markedly as we move from one sub-test to another, that is, the differences are not proportional to the number of items in the various sub-tests. Figure 2 illustrates the deviations more clearly (the percentage figures have been rounded off to the nearest whole numbers):

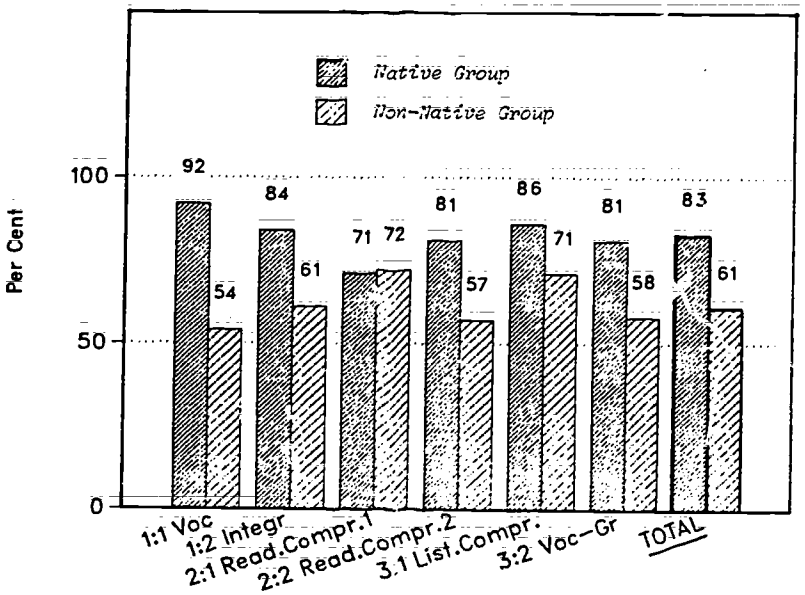


Figure 2 Main Test Results: Proportion of Correct Responses per Sub-Test

The most clear-cut difference is in the first sub-test (which measures word knowledge by means of multiple choice tasks). The English students here reach their highest score, while at the same time the Swedish students record their very lowest score. The result was very much the same in the York study and one may conclude that the sub-test in question has considerable discriminating power.

At the other extreme, showing no difference at all, or even a negative one seen from the English students' point of view, is sub-test 2:1 (which measures understanding of ordinary prose). The English students are far below their total test average (hitting their "low-water mark"), and the Swedish students are equally far above their average level (reaching their highest

score). This remarkable outcome is at variance with the corresponding result in York (where a sizable difference was obtained; cf Table 1). Reliability is relatively low (more so in the native group). It should be noted that restriction of range (cf explanation of this concept in Section 8.1) is not, in this case, a serious problem in the native group, the mean score being as low as 70.5% of the maximal score. (In other cases, low reliability indices may be explained by too high mean scores, preventing a natural distribution of individual scores.) A check of the results on individual items showed that the point biserial correlations (cf section 8.1) are relatively low; they are, for instance, lower than in sub-test 2:2, although this latter test is probably negatively affected by its higher mean. They are, furthermore, noticeably lower in the native group than in the non-native group.

Since it might be suspected that fatigue or boredom may have played a part in the weak native performance on the long Reading Test, a check was made of the average correct score frequency in the first vs. the second half of the test. The hypothesis was not borne out by the data. The average correct reponse rate was even higher in the later part of the test (67.4% vs. 73.4% in the two halves, respectively).

The inevitable conclusion is that the first section of the Reading Comprehenrion Test (2:1) did not function well in the native group (and not terribly well in the non-native group, either, judging by reliability and r_{pbis} figures). The possible reasons for this will be discussed in Chapter 11.

The remaining four sub-tests result in a fairly uniform pattern as far as mean score differences are concerned. The rank order, in terms of average correct response rates, is the same in the two groups and in the order 3:1 (Listening Comprehension Test), 1:2 (Integrative Test), 3:2 (Vocabulary-Grammar Test), 2:2 (Reading Comprehension Test, sentences). This measure of agreement may perhaps be taken as an indication of a certain homogeneity in the test. It may furthermore be noted

that part 2 of the Reading Test (i.e. sub-test 2:2) results in a very substantial difference, the second largest after the Vocabulary test (1:1), in sharp contrast to the outcome of the first part (i.e. sub-test 2:1). This is interesting but hardly surprising in view of the fact that sub-tests 1:1 and 2:2 are similar in form. Both are based on short snippets of text (often only one sentence) in which a word or a phrase has been deleted, and in both the testees choose the right answer among five options supplied. The main difference between the two is one of focus. In 1:1 the difficulty lies in the response part of each item (choosing among difficult words and phrases; the stimulus text presents no problem). In 2:2 the stimulus, the text, is the real test, whereas the options, in themselves, are unproblematic for the most part. The difference is not always apparent, however, and it would seem worthwhile to attempt a still clearer distinction between the two types of item. As it is now, sub-tests 1:1 and 2:2 probably tap much the same skills. The correlation coefficient for the relationship between results on the two tests is quite high (r=.69), taking into account the small number of items in 2:2, and this supports the hypothesis.

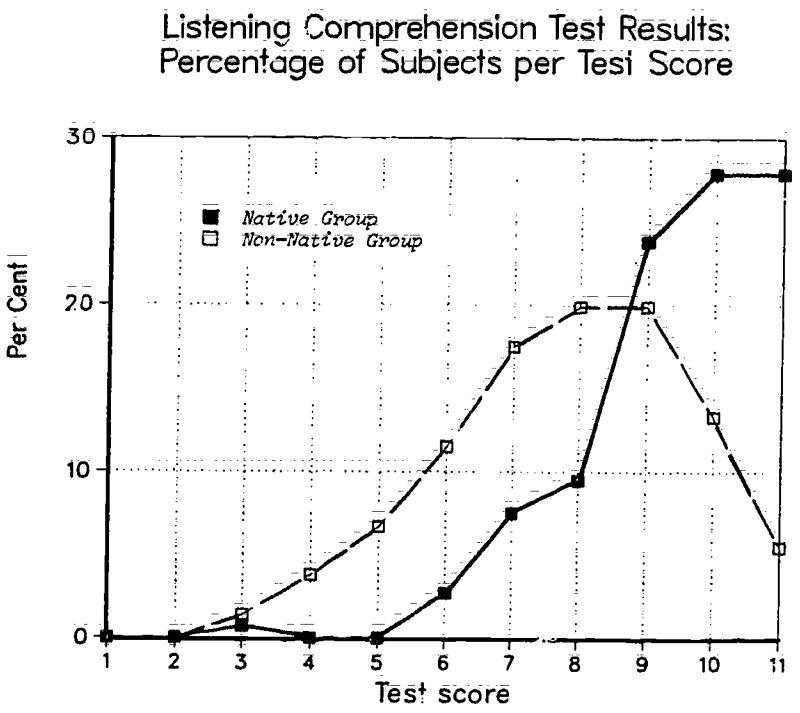
Sub-test 1:1 is, furthermore, cognate with sub-test 3:2, which also measures knowledge of vocabulary (in addition to grammar) and is based on very brief texts. (There is also a crucial difference between the two in that 1:1 consists of multiple choice tasks, whereas 3:2 consists of gap-filling tasks.) As in the previous experiment in York, the native scores are lower and the non-native scores higher in the latter sub-test, i.e. the Vocabulary-Grammar Test does not discriminate as well as the Vocabulary Test between native and non-native proficiency.

The results on the Vocabulary-Grammar Test will be analysed in more detail in Section 9.3 below.

The second smallest difference, i.e. after the exceptional case of the first Reading Test, is obtained on the Listening

Comprehension Test, where the English students are only some 15 % better than the Swedish students. It is fairly obvious that this figure does not reflect the "true" difference between the groups in general ability to understand spoken English (although it mirrors the difference on the particular tasks at hand). As figure 3 shows, the English and Swedish results overlap to a large extent:

Figure 3



Whereas the Listening Test thus proved to be of average difficulty for the Manchester students, it was more of a hurdle

to the native students in the York study (cf Table 1 in Section 3.1). The explanation for this lies, in all likelihood, in the tests themselves. It was felt among the test constructors that the 1983 version (used in York) was perhaps not a particularly good one; the 1985 version (used in Manchester) was more lively and also more authentic in that it involved the use of regionally coloured English. However, this latter version caused some harsh reactions from Swedish teachers, who complained (in the Questionnaire) that the test was unfair because, as they said (in summary) "the rate of speech was too high, voices were emotionally affected, and understanding was impeded by the dialect" (further details about attitudes will be given in Chapter 10). In 1983, there were very few complaints about the listening part of the test.

It is interesting to compare the results on the two listening tests against this background. There is compelling evidence that the students themselves did not find the 1985 version unduly difficult (in spite of the fact that it tested at a high level of comprehension). As Tables 1 and 2 show, both the Swedish and the British 1985 averages are actually up on the 1983 scores by 1 or 2 points. The Swedish level is raised from 64.3% to 70.6% of the maximal score and the native English level from 74.1% to 86.1% of the maximal score. The reliability indices are also slightly higher in the later version of the test.

In sum: the 1983 version of the test used relatively simple language, but the questions were relatively difficult. In the 1985 version it was the other way round.

From these figures it is difficult to draw any other conclusion than this: the 1985 version of sub-test 3:1 is technically a more appropriate and more valid measure of listening comprehension than the corresponding 1983 version. This does not in any way mean that we can ignore, or make light of, evaluative statements of the kind made in the Questionnaire. On the

contrary, they must always be taken into very careful consideration, when the various qualities of a test are finally weighed up.

The Integrative Test scores are very close to the total mean in both groups (which is not so surprising in view of the fact that the number of items in the sub-test makes up more than a third of the total number of items in the whole test; cf comments in Section 8.1). Reliability indices are, furthermore, very high for a sub-test, and although this, again, is partly explained by the relatively large number of items, there is reason to believe that the mode of assessing foreign language skills which the Integrative Test represents is a fairly dependable one.

More details on the Integrative Test results are given in Chapter 9.

8.3 Advanced native students' results

As was mentioned in Section 7.1, the native sample contained a sub-group of nine gifted "upper sixth form" students. They were 17-18 years old and were preparing for higher academic education. As these students' test results might be expected to differ from those obtained by the larger group, a separate analysis was made involving only this sub-sample of nine students. The results are set out in Table 3:

Table 3: Main Test Results: Advanced Native Group (N = 9)

Sub-Test (No. of items)	Mean score ($\bar{x}$)	Standard Deviation (s)	Mean Score (%)
1:1 Vocabulary (18)	17.56	0.72	97.5
1:2 Integrative (35)	33.67	1.00	96.2
2:1 Reading Compr 1 (12)	10.45	1.01	87.1
2:2 Reading Compr 2 (10)	9.67	0.71	96.7
3:1 Listening Compr (11)	10.56	0.73	96.0
3:2 Vocabulary-Gramm (14)	13.44	0.73	96.0
Total (100)	95.33	2.12	95.3

Overall, the results in this select group are some 12% higher than in the large group (of which the smaller group was a part). This may be taken as further proof that the National Test in English is indeed a test on which very proficient students obtain very high scores. The relationship between testees' knowledge of English and successful performance on the test is undoubtedly a very simple one: the better the English the higher the score. It should perhaps be emphasized again, at this point, that the test is primarily designed for establishing group means (and standard deviations) and that one should, therefore, be a little cautious when interpreting individual results (as well as results obtained by small groups). The individual score may actually deviate from the "true score" by several points, due to chance variation (cf explanation in Section 8.1.2).

Another feature that catches the eye in Table 3 is the result on the Reading Comprehension Test, part 1. Although the more advanced students manage to raise the proportion of correct responses considerably, we are still some way off the near-perfect target advocated by many testing experts (cf Chapter 4). All the other sub-tests are on a strikingly even level (cf Figure 4), and about 10% higher than the Reading Test (the long text):

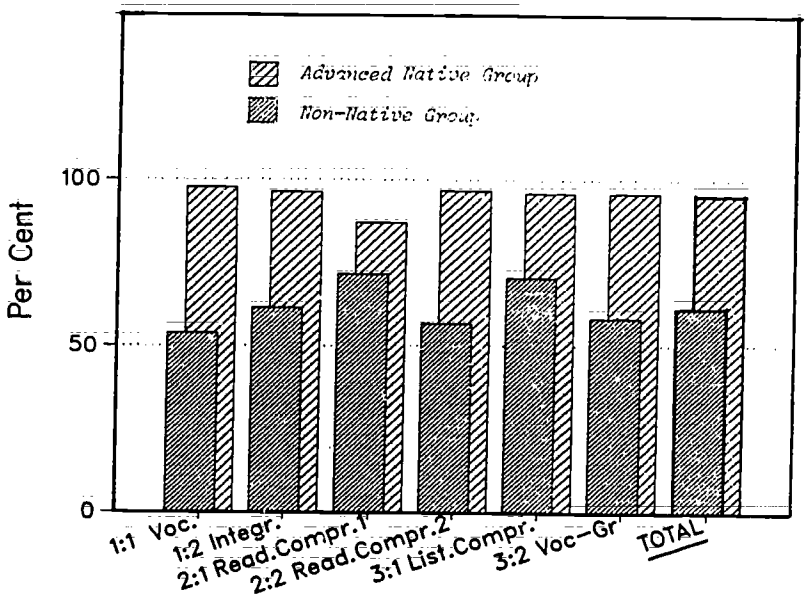


Figure 4 Main Test Results: Advanced Native Group in Comparison with Non-Native Group

Apparently, excellent reading skills, which must be taken for granted in this case, are not a sufficient condition for perfect scores on the present Reading Test. The question of what other abilities may be needed is discussed in Chapter 11.

It is interesting, furthermore, that the Listening Test, which has several features in common with the Reading Test (2:1), compares so well with other sub-tests. The story was quite different in the York study, as will be remembered (cf Table 1 in section 2.1); the York students were considerably less successful on the Listening Test than on the other sub-tests.

Further reference to the results obtained by the advanced group will be made, for control purposes, in the subsequent sections of the report.

9 THE RESULTS ON THE PRODUCTIVE PARTS OF THE TEST

9.1 Aims of analysis

As explained in Chapter 4, the experiment produced data which can be used for interpretation of the subjects' performance in absolute terms, in addition to evaluation in relative terms (which is the primary objective of the test). Absolute evaluation was particularly interesting in the national perspective, i.e. as a possible way of determining the quality of the "products" of foreign language instruction in Swedish schools. Moreover, concrete data on the characteristics of the English produced by groups of young native speakers today is of wide interest, not least to language testers and to practising teachers of English.

The more detailed qualitative analysis of actual language samples was made possible through the two sub-tests which require the students to formulate their own "answers" (the Integrative Test, 1:2, and the Grammar Vocabulary Test, 3:2) instead of choosing between given alternatives. The two tests will be analysed separately.

Since the Swedish sample of students comprised several thousand individuals, it was necessary to restrict the detailed analysis of the many varieties of answers to each item to a much smaller sub-sample, preferably one which was of approximately the same size as the English sample. To this end, a 1/200 sample was drawn, by random selection, from the original sample of upwards of 3,400 students. Thereby a more manageable group of 176 students was obtained. It should be noted that only random sampling procedures were used in order to arrive at this suitable number of subjects. Even though the group is extremely small in relation to the total number of students

taking the test (the proportion being 1:2,000), we may therefore safely assume that the group analysed faithfully represents the population of testees in every important respect. As pointed out in section 6.2, the large group of some 34,000 students who took the test is, furthermore, nearly identical with the entire age group of students in the 3- and 4-year lines of the Upper Secondary School, which makes the analysis all the more interesting.

9.2 Integrative Test results

9.2.1 Introductory notes

The Integrative Test (cf Appendix 2) measures a spectrum of abilities, and it is not always easy to single out and define each of these. However, there is little doubt that a substantial portion of general reading comprehension is normally needed in order to do the test successfully. (If the text is very easy in relation to the proficiency level of the students, reading comprehension will be a less decisive factor.) Other skills areas that are obviously directly involved are word knowledge (active as well as passive), control of grammar, facility in understanding and using idioms, and spelling. Both receptive and productive skills are required. The label Integrative is certainly well chosen for a test of this kind.

It is in the nature of things, however, that this type of test also allows quite drastic shifts of emphasis within the wide sphere of language ability indicated. It is possible, for instance, to stress the functional aspect by placing the blanks within certain types of set phrases and idioms like (21)"I was only _____(pulling) your leg" and (24)"It did cross my _____(mind)" or to put a premium on formal skills, e.g. grammatical accuracy, by concentrating on items such as (3)" ... it's like _____ (being) married to a sailor" or (9)"Are you thinking of _____(leaving) home, then?" Likewise, one may favour certain sub-areas within the major areas (e.g.

basic grammar in preference to advanced grammar etc). The present test could probably best be described as a balanced blend of functionally and structurally oriented tasks.

One of the native teachers who commented on the test expressed some concern over the idiomatic stamp of the language used: "Problems may arise with the Integrative Test because of idiomatic usages common to this area, and, possibly, age group. The students do not always seem to be acquainted with the idioms that clearly the sentence required for completion." Another teacher thought that the National Test as a whole was "More colloquial than expected with far greater use of idioms than in the teaching of foreign languages in England".

It should be noted that the actual words that are required in the gaps all belong to high-frequency bands (i.e. they are all very common, generally speaking). Thus the test is not, in spite of its appearance, a vocabulary test - at least not in any strict sense of the word. It should perhaps also be pointed out that many of the grammar points involved are not very advanced either.

The test is reproduced in Appendix 2.

9.2.2 Overall results

The results achieved by the two groups are set out in Table 4:

Table 4 Results per Item in the Integrative Test (1:2): Percentage of Correct Answers and Omissions in Native and Non-Native Groups (N = 134 and 172 respectively)

Item No.	Group Aver.(%) Nat.	Non-Nat.	Omissions Nat.	Non-Nat.	Item No.	Group Aver.(%) Nat.	Non-Nat.	Omissions Nat.	Non-Nat.
1	93.5	64.5	3.2	2.9	19	96.8	83.1	0.6	1.7
2	89.0	33.1	2.6	3.5	20	87.7	64.5	0	2.3
3	93.5	81.4	3.2	2.3	21	98.7	38.4	0.6	3.5
4	94.2	72.7	2.6	7.6	22	66.9	48.8	1.3	1.7
5	89.6	65.7	3.2	7.0	23	95.5	88.4	0.6	4.7
6	79.2	25.0	10.4	8.1	24	99.4	64.5	0.6	0
7	81.2	66.9	3.2	3.5	25	68.2	76.7	1.9	0
8	62.3	59.3	5.2	5.8	26	95.5	52.3	0	7.6
9	86.4	72.1	3.9	5.8	27	88.3	29.7	0.6	5.2
10	94.2	70.3	3.2	4.7	28	92.2	83.7	1.9	2.3
11	72.1	75.0	3.2	2.9	29	86.4	73.3	4.5	7.6
12	86.4	49.4	3.2	3.5	30	95.5	80.2	1.3	4.1
13	52.6	72.1	4.5	9.3	31	71.4	73.3	5.2	1.7
14	83.8	82.6	2.6	4.1	32	93.5	68.0	1.3	4.1
15	61.7	62.8	5.2	4.7	33	81.8	59.3	1.9	4.7
16	94.8	89.5	0	0.6	34	68.2	47.1	1.3	5.2
17	79.9	52.9	0.6	5.2	35	85.7	76.7	1.3	1.2
18	86.4	54.1	0	0.6	Total	83.5	64.5	2.4	4.0

The percentage of omitted answers is not very high (on average some 3%). Only in one case does it reach the 10% level (on item 6, which obviously baffled both native and non-native students). The proportion of omissions is, however, much smaller in the other productive sub-test, the Vocabulary-Grammar Test (3:2; cf Section 9.3). This difference probably shows that the students found the tasks in the Integrative Test somewhat more confusing, or difficult to respond to, than those in the Vocabulary-Grammar Test.

Close inspection of the figures shows that there is very little, if any, correspondence between native and non-native results. (Calculation of the so-called Pearson product-moment correlation coefficient yields a value of r = .13; Spearman's rank correlation coefficient is .23) The following scattergram (Figure 5) illustrates this lack of agreement:

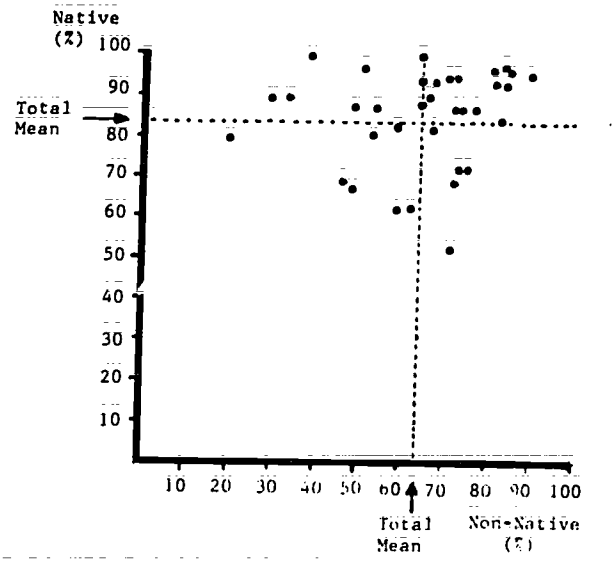


Figure 5: The Relationship between Native and Non-native Item Mean Percentages on the Integrative Test (1:2)

In about a third of the cases, the values depart from group averages in opposite directions in the two groups (i.e. the percentage of correct answers is higher than average in the native group and lower than average in the non-native group, or vice versa). These data suggest that native and non-native students often experienced quite different problems when taking this sub-test. The matter will be further elucidated in Sections 9.2.3 - 9.2.6, which identify the linguistic areas in which the difference between native and non-native competence stands out most clearly.

Another (familiar) trend is for non-native students to show much greater variation in ability than native students. The non-native facility indexes, i.e. the proportions of correct scores (henceforth "pass rates"), range from 25.0% on item 6 to 89.5% on item 16 (i.e. the span covered two thirds of the whole range). The corresponding native figures are 52.6% (on item 13) and 99.4 (on item 24). In this case, the span covers

the upper half of the scale; the observations tend to cluster within the 80-95% range.

These figures tell us, among other things, that native competence is relatively homogeneous in the areas covered by the test (which was expected) and that non-native competence, as demonstrated by the present sample of Swedish students, is characterized by great variation depending on what particular problems we are dealing with.

The widest gaps between native and non-native performance levels are to be found on items 21, 27, 2, and 6 (in descending order of magnitude, all exceeding a difference of 50%). These four items will be examined individually below.

9.2.3 Item 21: 'to pull someone's leg'

Item 21 tested the students' knowledge of the idiom 'pull someone's leg', a very common figure of speech which means 'to make fun of a person in a playful way'. The sentence in which it occurred runs as follows:

PAM: (laughs) Come on, Ken, I was only _____ your leg.

Only one native speaker out of 154 got it wrong and supplied 'putting' (which obviously was a sheer slip of the pen and not really a linguistic mistake). Undoubtedly, the expression is very well known by competent users of the language and it belongs to the natural repertoire of linguistic forms which all adult native speakers possess (cf Irujo, 1986:288, who investigated transfer in the acquisition of idioms, among others 'pull somebody's leg'.)

The non-native students were definitely inferior to the native students in respect of idiomatic command of the language, as their much lower pass rate (38.4%) on the same task shows (cf

also the pronounced differences on items 26 and 24, which both involve idiomatic expressions). Only one in three (or a little more) was familiar with the phrase and could insert the missing word in its correct form. On the other hand, almost everyone (94%) had control of the grammar involved, i.e. used a present participle form of a verb (an ing-form) in their answers. (Many a brave attempt at a lucky shot score was made, e.g. 'I was only /biting, breaking, crossing, kicking, twisting, scratching, testing .../ your leg".)

9.2.4 Item 27: 'as if'

A very pronounced difference between scores was also obtained on item 27, which required insertion of the subordinating conjunction 'as' in the following question:

KEN: Well, it's not _____ if we were old pals, is it?

The missing word is, thus, part of a subordinating phrase (a so-called compound subordinator).

The task did not pose a problem in the native group (where above average performance was recorded), whereas it was a real stumbling-block to the other group. Only 30% of the Swedish students answered it correctly. The main problem in this group was whether or not 'like' could be used (it was chosen by 32% of the students). The reason is that both 'as' and 'like' may correspond to one and the same word in Swedish, viz. 'som', which can either be used as a conjunction in adverbial clauses of comparison (corresponding to 'as') or as a preposition expressing comparison in a prepositional phrase (corresponding to 'like'; for further explanation and examples, see Svartvik and Sager 1983:336). In about a third of the cases, the Swedish students failed to realize that it was a subordinator that was missing.

The main lesson to be learnt from the outcome of this item is that Swedish students have great difficulty in handling this particular instance of clause connection. In the main, the problem arises from inability to make the right distinction between the words 'as' (used as a conjunction) and 'like' (used in a prepositional function). A contributing factor may be the possibility of using 'like' in place of 'as if' in very informal language, particularly in American English. (The combination 'like if' is not possible, of course.)

9.2.5 Item 2: 'hardly ever'

Item 2 was another poser; it was only solved by a third of the students in the Swedish group:

> PAM: ... His job takes him all over the country, and abroad too, sometimes, so he's hardly _____ at home.

The single word that best fits into this frame is 'ever', and this was chosen by the vast majority (86%) of the English students. (Other rather less appropriate but acceptable suggestions were 'living' and 'seen'.) The adjunct 'hardly ever' expresses a distinct time concept and it has wide applicability in that it is not restricted to any particular linguistic register or to any particular mode of language use. In view of this, it is hardly surprising that the native students passed the item with flying colours. (The few unaccepted replies disclosed a different, and not entirely illogical, train of thought.)

What is surprising, however, is the fact that this very frequent adverb phrase was such a hard nut to crack in the Swedish group. The answers showed that the students were not in doubt as to what concept or notion was implied (i.e. comprehension of the text was not a problem). Thus 39% of those who failed put down 'never' in the gap, others suggested 'rarely'

and 'seldom'. Some of the answers, but fewer than expected (3-4%), reflected influence from the Swedish language ('any', 'anything' etc).

Thus item 2 (testing the time adverb 'hardly ever') seems to have revealed a weak spot in Swedish students' command of English. In practice, this "lacuna" is probably filled by use of the logical equivalent 'almost never'.

9 ?.6 Item 6: 'when he <u>does</u> turn up'

The last task that exhibited a very large difference between English and Swedish pass rates was item 6. The context was this:

PAM: ... You never have a chance to get fed _____ (<u>up</u>) with a husband who's only at home occasionally. They're like a couple of proper lovebirds when he _____ turn up.

The only possible completions are, conceivably, 'does' and 'can'. The task requires close reading of the text. Adverbs like 'eventually', 'suddenly', and 'sometimes' (suggested by 4% of the native students and 33% of the non-native students) are of course incorrect here because of the infinitive following the blank. The crucial element tested is the emphatic use of the auxiliary (i.e. 'do' in the first place).

It should be noted that the item turned out to be on the difficult side in both groups (extremely difficult in the non-native group; a little below average in the native group). This is reflected not only in the small number of correct responses, but also in the large number of omissions. Evidently, some of the students got confused at the task and did not know how to respond; in the native group this is the main explanation of the poor result.

In the Swedish group, the pass rate was exceptionally low (lower than on the very difficult 'as if' problem discussed above) and this is of course the most remarkable observation in connection with item 6. The low score can hardly be due to lack of understanding of the text surrounding the gap (i.e. poor reading comprehension, generally), nor can it be attributed to difficult vocabulary or unfamiliar idiomatic language (as in the case of item 21). The direct cause may therefore be inability to recognize a syntactic pattern which typically fits in with the use of the auxiliary 'do' for emphasis. Another plausible explanation may in fact be lack of ingenuity. It goes without saying that the task does stretch one's power of imagination a little. It should furthermore be noted that the task is different from the other tasks in the test in that the missing word is lexically empty (cf for instance Crystal 1985:108); it only has grammatical function.

The analysis of results carried out so far suggests that Swedish students are, in comparison with native students, fairly weak as far as some quite fundamental points of grammar are concerned. (Of course knowledge of grammar is only one ingredient in the skill it takes to complete the tasks successfully, as pointed out earlier; it seems to be the key ability, however, judging by the types of error committed.) Not surprisingly, they are also very much weaker than native students in the area of idiomatic use of the language.

9.2.7 "Reversed" results

As a contrast to the foregoing analysis, it may be interesting to look at items that resulted in a reversed difference, i.e. items on which the non-native students were actually ahead of the native students. The most characteristic of these are Nos. 13, 25, and 11.

The reversed pattern of results (favouring the Swedish students) involved one item (No. 13) where nearly a third of the

native students misread the text (or did not read it closely enough). Instead of inserting 'too' in the frame '... if it isn't _____ personal a question', they produced 'a', which showed that they did not notice that the article was already there. The Swedish students mastered the difficulty (essentially a word order problem) quite well.

Both of the remaining items (25 and 11) require attention to - as well as active control of - tag questions, No. 25 directly and No. 11 indirectly. Again the Swedish students were quite adept at supplying acceptable answers. The non-native students were rather less successful; in 25, this was due to inconsistent or inappropriate choice of tense (present instead of past) and in 11, choice of the wrong verb ('shouldn't' instead of 'won't').

The fact that we get a small number of reversed scores is by no means sensational. They may actually constitute more or less fortuitous outcomes, considering the number of factors that influence the totality of responses in a relatively comprehensive test battery (cf discussion in Section 10.3.1). There is also the possibility - or even the likelihood - that they are, in part at least, an effect of the type of referencing made when the test was developed (i.e. referencing against a non-native population). As pointed out by Oller (1979):

> If the variance in the performance of natives is not completely similar to the variance in the performance of non-natives, it follows that items which work well in relation to the variance in one will not necessarily work well in relation to the variance on the other. In fact, we should predict that some of the items that are easy for native speakers should be difficult for non-natives and vice versa. .... some of the items in the test will tend to gravitate toward portions of variance in the reference population that are not characteristic of normal language use by native speakers. Hence, some of the items on a test referenced against non-native performance will be more difficult for natives than for non-natives, and many of the items on such tests may have little or nothing to do with actual ability to communicate in the tested language. (p 201f)

It would seem that the above contention relates, to the extent it is correct, more to content validity than to construct validity (cf Chapter 2). A test which measures the attainment of certain given skills ("a content" defined by some specific criteria), rather than a hypothesized general ability (a construct), is probably more likely to contain some items which result in aberrant native scores.

9.2.8 Conclusions

Our analysis of Integrative Test results may be summed up as follows:

There was hardly any correlation between native and non-native performance at the item level. High and low achievements did not coincide systematically. Items that were easy in the native group were often difficult in the non-native group and vice versa. This may probably be taken to mean that the typical Swedish student has developed a structure of skills which differs from that of the native English speaker. The reason for this is obviously the fact that the Swedish student has learnt the language in a more or less artificial situation (rather than in a natural language learning environment). Thus, it might in fact be argued that the ability of our non-native students to answer certain Integrative Test items correctly is only loosely related to the kind of ability native speakers display when they use the language in a "normal" linguistic situation. It is probably inevitable that this should be so, at least to a certain extent, given the conditions for foreign language learning in schools, but the observation merits close attention.

Item pass rates were, furthermore, widely variable in the Swedish group (and much less so in the English group). Areas which seemed to cause problems were idiomatic usage and certain grammatical structures (such as the subordinator 'as if', the time adverb 'hardly ever', and emphatic 'do').

There is little evidence that Swedish students did not, by and large, comprehend the text on which the items were based. (The cloze testing technique, of which the Integrative Test is a variety, was originally devised as a method for measuring readability; cf Taylor 1953.) The foremost hindrance to higher results was a certain lack of formal language skills. In other words, what the present Integrative Test seems to be measuring, above all, is knowledge of language forms, i.e. words (mostly very common ones), phrases, grammar, spelling etc. If, on the other hand, the test had been based on a text of a slightly higher level of difficulty, the element of reading comprehension would probably have been a more crucial determinant of test scores.

9.3 Vocabulary-Grammar Test results

9.3.1 Introductory notes

The Vocabulary-Grammar Test (cf Appendix 3) measures, in keeping with its designation, word knowledge and mastery of grammar, but it also measures, to a limited degree, more functional linguistic abilities. One is the ability to handle and express "ideational content" or, in more topical terminology, general language notions (e.g. 'existence/ non-existence', 'possibility/ impossibility'), another is the ability to use the language for a purpose or, in recent vernacular, in order to perform language functions (e.g. 'make suggestions', 'ask permission'). (For a practical exposition of the import and nature of notional-functional categories in language learning, see van Ek and Alexander, 1980). The role of formal accuracy is further played down in this sub-test in that spelling errors are disregarded completely (even "grammatical" spelling errors such as 'comeing'; cf Section 5.3). In a way, the name of the test is, therefore, somewhat of a misnomer; at any rate, it does not capture the essence of all the skills it takes to solve the tasks successfully.

It should also be noted that the type of task used imposes certain constraints on what features of grammar and lexis one may measure, as the sample in Table 4 makes abundantly clear. Given that one wants to test (as in 3:2) not only individual words but successions of words connected with each other (minimally two-word strings), it proves to be very difficult, to begin with, not to involve in the task the verbal part of the clause or sentence in which the blank appears. That is to say, the frame that surrounds the blank may easily be modelled to trigger various finite (sometimes non-finite) verb constructions; it is far more difficult to see to it that the frame requires the student to use, for instance, a complex noun phrase such as a verbless modifier + noun construction (on the pattern 'a terribly important meeting'). Other areas which seem hard to get at by the open multiple-word gap technique are word order and use of adverbs as modifiers. Similarly, at the lexical level, it is very difficult to elicit anything else than quite trivial high-frequency vocabulary; taxing the student's ability to produce less common words and phrases is next to impossible.

All this may not seem to be a serious disadvantage, at least not in the context of norm-referenced testing of proficiency, but we should be aware of the fact that it may perhaps have wider implications; theoretically it may lead to a certain bias in the priorities the teacher makes in his instruction of students. The limitations should of course also be kept in mind when we interpret the outcome of our study.

9.3.2 Overall results

With these remarks in mind, we will now proceed to a scrutiny of the results "item-wise". In Table 4 an attempt is made to characterize the items on the basis of their most prominent features, which, in spite of what was said earlier, are of a grammatical kind in most cases. For practical reasons we will restrict the identification to this aspect of the tasks. It

should also be mentioned that there are sometimes ways around the grammatical obstacles specified. Particularly the native speakers delivered, on and off, unpredicted correct answers which did not require use of the grammar anticipated. However, the overwhelming majority of responses did involve the structures listed in the table.

Very few answers were missed out (13 out of a maximum of 2,170 in the native group, and 13 out of 2,464 in the non-native group, i.e. there was no response in roughly 0.5% of the cases). It was therefore judged unnecessary to include the frequencies of incorrect responses; they simply make up the remaining percentages up to 100%.

Table 5 Results per Item in the Grammar-Vocabulary Test (3:2): Percentage of Correct Answers in Native and Non-Native Groups

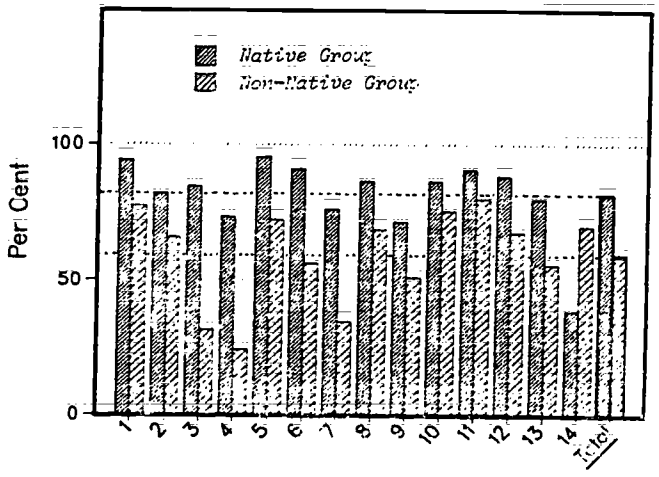
Item No.	Core Content	Mean Score (%) Native Group (N=155)	Non-Native Group (N=176)
1	verb phrase (present tense) realized as a 'do' or passive constr.	94.2	77.3
2	verb phrase (past perfect) + obj.	81.9	65.9
3	verb phrase (future tense) + idiom 'make a mistake' + def.art.	84.5	31.8
4	existential 'there' + idiom 'no point in'	73.5	20.4
5	verb phrase ('would' + infin.) in conditional sentence	95.5	72.2
6	verb phrase (pr.tense) with 'want', 'allow' etc + non-finite clause	91.0	56.2
7	non-finite verb construction using idiom 'change one's mind'	76.1	34.7
8	verb phrase ('would','could' etc + perf.inf.) in conditional sent.	86.5	68.7
9	non-finite verb construction after 'hate' + idiom 'look after'	71.6	51.1
10	existential 'there'/'it' as prop subject + verb phrase (fut.tense)	86.5	75.6
11	idiom 'let's go to ...' (making suggestion)	91.0	80.1
12	interrogative clause construction	88.4	67.6
13	the superlative of adj. + verb phrase (pres. perfect tense)	80.0	55.7
14	idiom (polite phrase) 'do you mind if I ...'	38.7	69.9
	T o t a l	81.4	59.4

Before going into a discussion of the results, we will make some comments on the contents of the test as reflected in the specification above.

As the table shows, the native students perform on a fairly high and even level throughout the test (except for the spectacular "nosedive" on the last item, which will be discussed later). The Swedish students' performance is of course generally less accurate and is much more irregular. The rank correlation between native and non-native item averages is .57, indicating only very moderate agreement between the two sets of scores.

Figure 6 illustrat s the pattern of results more clearly:

Figure 6
Results per Item in the Vocabulary—Grammar Test



As could be expected, the English students made very few purely grammatical or lexical mistakes, although they did produce a certain amount of careless or sloppy language which had to be marked down as formally incorrect in the contexts

provided. By far the most common reason for lost points in the English group was lack of attention to detail in the texts and misunderstanding of the prompts. In the Swedish sample most mistakes were grammatical.

The lowest correct scores in the Swedish group were recorded on items 4, 3, and 7. As they were also the ones which resulted in the widest gaps between native and non-native proficiency (cf the corresponding analysis of Integrative Test results in Section 9.2.3-9.2.6), we will start by looking at each of these in turn.

9.3.3 Item 4: 'there's no point (in worrying)...

Item 4 consisted of the following exchange:

LEN: The damage is done
and _____
in worrying about the consequences now.
RON: That's easy for you to say.

There were about a hundred different answers in the Swedish group; only about 6% of the students (39% in the English group) produced 'there's no (little) point', which was the correct answer anticipated in the key. Accepted variants in the Swedish group were, for instance, 'there's no sense (use)' and 'I'm not interested'. The preposition 'in' after the gap was crucial and prevented acceptance of the phrase 'it's no good (use)', which was suggested by 10% of the native speakers and by 16% of the non-native speakers. (Incidentally, there was no instance of 'it's no good' in the latter group, whereas 'no good' and 'no use' were equally common in the native group.) Similarly, nine native students (6%) and two non-native students (1%) used the phrase 'there's no need'. Other answers that might have been accepted, had the preposition not been present, were 'I (we, you) can stop', 'you shouldn't

begin', 'don't start', 'it is too late', ' don't waste your time' (all in the Swedish group).

The fact that attention to a small detail in the text (presence of the preposition 'in') had a noticeable effect on the results is a little infelicitous no doubt. At least from a broad communicative point of view the incorrect examples quoted above bear witness to good functional control of the language. It would seem to be quite important to try to avoid niceties of this kind in the construction of multiple-word gap tasks.

However, the major problem underlying the "sub-standard" Swedish performance was the students' general inability to cope with the problem at hand, in a formal as well as in a functional perspective. The following list of unsuccessful attempted answers illustrates the nature of this overriding problem: 'blame yourself', 'no not go', 'don't be to much', 'don't stay', 'it doesn't be better', 'it now (!) idear to be', , 'now you have begin', 'that's no idea', 'there is no reson', 'why stay', 'you do wrong' 'it's no idea', 'it's no matter', 'it's no need', 'it's no sense (sence)' etc. Although the students had a lot of trouble getting the idiom right, the main source of error was the choice of subject, i.e. deciding on which of the two pronouns 'it' and 'there' to use. They occurred in altogether 65% of the answers (31% and 34%, respectively). The percentage of correct applications (involving 'there') was only 23% (as against 74% in the native group). However, the snag discussed above (caused by the presence of the preposition 'in') should be kept in mind. Some of the students who misused 'it' here may actually have mastered it in a different situation.

Further comments on the students' way of handling the 'it/ 'there' problem will be given in Section 9.3.9 (item 10).

9.3.4 Item 3: 'he won't make the (same mistake)...'

The next task to be discussed, Item 3, was based on the following two sentences:

> It's quite clear that Tom messed up the deal, but he's learnt his lesson by now. I'm sure _____ _____ same mistake again.

This item is very "tight" and does not allow the same variety of responses as the previous one (No. 4). The context necessitates use of the pronoun 'he' as subject + the negative of the verb 'make' + a future tense construction (use of the present tense, as in Swedish, will not do) + the definite article (which cannot be left out). There are, thus, several discrete grammar points involved. Furthermore, the item is semantically unambiguous, which means that there can be no abstruse interpretations of the situation that might be hard to evaluate; the answers showed very clearly that all students knew exactly what the task required of them in terms of language. For these reasons, it is interesting to look at the results in detail.

The subject ('he') was missed out in a couple of cases and no more; thus this point was no problem in either group.

The first real problem was the choice of the verb. The Swedish students were very uncertain and opted for 'do' in 56% of the cases. 'Make', which is the only universally accepted verb in this context, was chosen by 42% of the students.

The predominant choice in the native group was of course 'make'; 89% of the English students used this verb. Interestingly enough, though, 'do' was chosen by as many as 9% of the students (it also occurred in the advanced group). This piece of evidence indicates, possibly, that we are here dealing with an instance of what is sometimes referred to as 'divided usage'. (Language change may sometimes occur very rapidly in the young generation; cf for instance the research carried out

by Peter Trudgill, the sociolinguist; Chambers and Trudgil), 1980.) On the other hand, the majority of native speakers would no doubt regard 'do a mistake' as a typical example of sub-standard or sloppy language.

In the analysis of the students' choice of tense, we have treated 'wan't', 'want', and 'woun't' as spelling mistakes (for 'won't') and not as grammatical mistakes. (As will be remembered, spelling errors were not penalized in this subtest.)

The choice of the right tense (the future) was made as a matter of course in the native group. Without exception the students used a proper marker for future in their answers.

In Swedish, the present tense is frequently used to indicate future time, not only in conditional and temporal clauses (as in English), but in (nominal) 'that'-clauses as well, and it might therefore have been suspected that the inappropriate use of present tense forms would have been a major type of mistake in the Swedish group. This was not the case. Only 9% of the students used present tense verb phrases like 'doesn't make'. (On the other hand, the answers these students produced were often grotesquely wrong, e.g. 'he don't do', 'he don't want do the', 'he not do the', 'af that he's made the'.) Nine students in ten (91%) correctly used a form of 'will' or 'be going to'. Six students in this group (3%) preferred to use 'would' (as a 'mood marker', hypothetical 'would'; cf Quirk et al 1985, section 4.64), which of course must be accepted, too. ('Would' was likewise chosen by 3% of the students in the native group.)

The definite article, finally, is potentially a problem to Swedes in the context in which it occurs here, simply because the indefinite form of the noun is used in the corresponding structure in Swedish. However, Swedish students at this level seem to have internalized the relevant rule quite well. The article was absent in some 12% of the papers; it may

furthermore be assumed that it had been left out by mistake in some cases.

The main reason for the very low Swedish score on item 3 was thus unfamiliarity with the idiom 'make a mistake', or more precisely, uncertainty as to what verb to use in this idiom. Choice of tense and use of the definite article were only minor problems.

9.3.5 Item 7: '(make) him change his (mind)...'

Item 7 consisted of another mini-dialogue:

DAVE: It's irritating that the boss refuses to discuss our project.

ALAN: Yes, isn't it? I've tried to make _____ mind a couple of times, but he won't.

The task involves manipulation of the expression 'change one's mind' within the confines of a non-finite clause construction ('to make somebody do something'). It must be regarded as a fairly complex grammatical structure, and there is the added complication that the slot happens to evoke a competing image involving the phrase 'make up one's mind'. This, of course, makes the task less transparent; the correct solution does not come to mind very easily.

Half the students in the Swedish group (48%) sensed that a non-finite construction was unavoidable, but only 34% used the logically apposite as well as grammatically correct sequence 'him change his'. 9% tried the less appropriate (although structurally possible) phrase 'him make up his'; 5% had a go at non-finite constructions that were impossible on both counts (e.g. 'him made up his', 'him getting it off his', 'he changing his').

All the remaining attempts to find a suitable answer were futile. The pull of the competing structure ('make up one's mind') proved to be very strong indeed. No less than 40% of the students entered 'up his' in the gap; one cannot help feeling that some of these students were actually lured into a trap here. Nevertheless, they created a sentence which makes no sense at all.

The native students used non-finite clause constructions to a much higher degree. Three out of four (i.e. 75%) wrote precisely 'him change his'; two students produced the equally acceptable 'him speak his'. A few (5%), including one student in the advanced group, chose 'him make up his' and 10% put down 'up his' (which again shows that this item tends to catch students unawares). Then there were some oddities like 'them change their', 'him understand' (this student probably did not notice the word 'mind' at the end of the line), 'take it of his', 'his change his' (obviously a slip of the pen).

9.3.6 Final note on items 4, 3 and 7

Overall, the Swedish students did very poorly on the items now discussed (Nos. 4, 3, and 7). The reason for this was the combined effect of shaky control of the structural patterns involved and insufficient familiarity with the words and phrases required for the completion of the tasks. It should be noted that the relevant idioms (i.e. 'no point/no use in', 'make a mistake', and 'change one's mind') play a very important role in each of the three items. Furthermore, they are all very common in everyday spoken (and written) English; this is shown by the fact that the native students did not have more trouble with these particular items than with the rest of the items, generally speaking. Their average pass rate was 78% (to be compared with the average of the rest of the items, which was 82%). In the Swedish group, the pass rate was considerably lower than average, viz. 30% (the average of the rest of the items was 67%). The net result of the analysis may therefore be that Swedish students' acquaintance with English idioms

does not quite match their level of general grammatical mastery of the language. The evidence pointed in the same direction in the Integrative Test (cf Section 9.2).

Against the background of the point just made, it is interesting to try to determine areas in which Swedish students tend to excel (relatively speaking, that is). This may be done by investigating items on which pass rates were particularly high in the Swedish group, i.e. Nos. 11, 1, and 10 (in descending order of facility).

9.3.7 Item 11: 'let's go to a restaurant'

Item 11 tested a typical language function exponent, namely 'let's go to a (restaurant)' (the language function being 'making a suggestion'):

LINDA: What a victory! We ought to celebrate.
BRIAN: Yes, _____
restaurant for a really good meal.
LINDA: That's an excellent idea.

The outstanding characteristic of the results was that there were very few gross mistakes; likewise, there was no sign of misinterpretation of the task in either group. Furthermore, the grammatical component is only marginally present (the most obvious completion being a formulaic type of expression and little more) and it may therefore be very telling to compare English and Swedish performance levels as far as this particular function of the language is concerned.

Starting with the English group (who were 91% correct), we may simply conclude that the majority of zero scores arose out of carelessness. Students did not pay close enough attention to the whole of the latter part of the conversation (the segment after the blank) as the following sample answers show: 'I know a good', 'I know just the', 'and go to a really good', 'let's

go for a meal'. In some answers there was an ellipsis of the verb, resulting in word strings such as 'at a', 'in a', 'in an excellent'.

The pattern was, actually, very much the same in the Swedish group, i.e. most students (80%) produced a proper formula for suggesting an action ('let me take you to a', 'let's go to a', 'let's find a', 'we could go to a', 'why don't we go to a' etc) and a number of them failed to take the latter part of the dialogue into due account, thereby coming up with answers like 'in a', 'I know a good' etc.

The performance of the two groups differed mainly in two respects. Firstly, the natives who scored their points were in fact "more" correct, most of the time, than the non-natives who managed to "pass", i.e. there was a much larger number of acceptable (rather than perfect) answers in the latter group (e.g. 'we should go to a', 'let us go out to a', 'shall we go to a', 'would you like to go to a'). Secondly, the incorrect responses in the native group were nearly all discarded on conceptual (not formal) grounds, whereas the incorrect answers in the non-native group contained a considerable amount of formal mistakes in addition to being generally misconceived (e.g. 'let's visite (vissit)', 'visiting a good', 'we are going to the', 'we go out to', 'we would go to a').

The conclusion is that, although the Swedish students demonstrated sound communicative ability on this task, there is a much wider gap between the two groups than the actual figures may lead one to believe. In other words, the testing technique utilized here is not particularly sensitive to real differences in language skills. The key problem resides of course, as ever (at least in a test of this kind), in the evaluative criteria employed. If a more elaborate and more comprehensive marking system could be worked out, this would probably improve the potential of the Grammar-Vocabulary Test considerably (while at the same time add to the teachers' burden of marking, which is an obvious disadvantage).

9.3.0 Item 1: 'What do they mean by that?'

The first item on our list of Swedish "success tasks", item 1, reads as follows:

JIM: This advertisement says that
the machine is "fool-proof".
What _____ by that,
Daddy?
DAD: That it's so simple that anybody can handle
it,even a fool.

The task resulted in the second highest facility value, in both groups, and it obviously served, in some measure, as a warm-up item. There was very little variation among the answers, which was natural since no more than two words were needed, minimally, in the gap. Nobody seemed to be hesitant as to what to write (only how to write it).

A prerequisite condition to the solution of this particular language problem is the ability to produce a grammatically correct verb phrase which fits into the slot (either in the form of a 'do' construction or in the form of a passive construction). Another requirement is active mastery of the lexical item 'mean'.

Only one out of four students in the Swedish group was unable to measure up to this not too daunting challenge. The relatively few failures were mostly due to weak grammar (everybody knew 'mean', or an equivalent) as the following examples show: 'are they meaning', 'is meen', 'does it means', 'does they mean (meane)', 'does they meant'. Faulty 'do' construction, as exemplified here, occurred in 9% of the cases and this fact tends to spoil the fairly bright picture; after all, the ability to form a straightforward present tense verb phrase using 'do' must be considered very basic by any standard.

The native errors, produced by 6% of the students, were mostly very trivial; four students used the wrong referential pronoun ('he' instead of 'they'), two students omitted parts of the required insertion (resulting in the truncated answers 'do they' and 'do you think'), one student used the past tense instead of the present, and two students -- lo and behold! -- made a mess of their 'do' constructions, producing 'do you meant' and 'does he means' (so, EFL teachers, don't give up just yet ...).

9.3.9 Item 10: 'it will rain'

Lastly item 10, which ranked third on the list of the most positive achievements in the Swedish group:

ROBIN: The forecast says that _____ all over the country tomorrow and probably for the rest of the week.
OSCAR: Oh dear, more rain! I was hoping for some sunshine for a change.

In order to be successful in this task t testee will normally need to have command of three discrete linguistic elements at the same time, namely existential 'there' as subject (alternatively 'it' as empty 'prop' subject), future time reference in the verb, and some fitting lexical item for the notion of 'bad weather' (such as 'rain', 'showers', 'wet' etc).

Three out of four students in the Swedish group proved that they were adequately equipped for this multiple task, which must be considered a fairly gratifying achievement.

There was, however, among those who failed, a great deal of confusion as to the use of 'there' (which must be followed by an indefinite noun phrase as the 'notional subject'; cf for

instance Quirk et al, 1985, section 19.45) and 'it' (used as a 'prop' subject in expressions denoting atmospheric conditions; cf ibid, section 10.26). This difficulty is well known to all teachers of English in Sweden; it derives from the fact that one and the same pronoun in Swedish ('det') may perform both of the above functions. (In English 'there' and 'it' are of course not interchangeable.) Typical mistakes were 'it will be rain' and 'it's going to be bad weather'. Using 'it' in p of 'there' was the predominant type of error (it was present in 14% of the cases); 'there' was wrongly used only once, actually ('there will be raining').

The lesson to be learnt here is that many Swedish students have not yet, at their present stage of learning, developed a proper feel for the use of 'it' as a 'prop' subject and also that their use of existential 'there' is relatively scarce when an 'it' subject construction may equally well be substituted. The following table verifies this latter point (notice that the figures have been calculated on the basis of correct responses only):

Table 6 Distribution of Correct Responses over Answer Types in Native and Non-Native Groups (Sub-test 3:2, Item 10)

Answer Type	Correct Responses (%)	
	Native	Non-Native
(a) 'it will (it'll) rain /be raining .../'	56	81
(b) 'it is (it's) going to rain /be raining/'	17	9
(c) 'there will (there'll) be rain /showers, storms/'	27	10

As the table shows, expres s using 'there' were almost three times as common in the native group. The use of 'there' thus seems to be markedly underrepresented among Swedish studer , at least in relation to its use in our experimental native sample. The impression is strongly reinforced by the outcome of item 4 (described earlier). In that task, only 23%

of the Swedish students produced an adequate existential 'there' construction, as against 74% of the native students.

Comparing answer types (a) and (b), we may further conclude that the use of the phrase 'be going to' as a means of expressing future time (with an 'impersonal' subject) is less common among Swedish students than among native students. It was used twice as often in the latter group. The alternative, 'will', was preferred by 8 students out of 10 in the Swedish group. The relationship was only 5 or 6 to 10 in the English group.

In sum, the analysis of the results on item 10 showed that most Swedish students (of the category represented in this study) are able to deal with the language requirements involved in this particular task. The main problem identified pertains to a certain inability to use 'it' as 'prop' subject in the right context. It was often used in cases where existential 'there' would have been the appropriate choice; 'there', in turn, was used much less frequently in the Swedish group than in the English group. Finally, a tendency to "overuse" 'will' (relative to 'be going to') as a marker for future time was noted.

9.3.10 Item 14: 'Do you mind if I...'

Before summing up the results presented in this section of the report, we will say a few words about the very last task in the test, item 14, which resulted in a downright demise of the usual prowess shown by the native students. The wording of the task was this:

SECRETARY: I've got to pick up my child at the nursery today. _____ leave early?
MANAGER: No, that's all right. Have a nice weekend.
SECRETARY: Thanks. You too.

The outcome was a complete reverse of what was found in the rest of the test; the native pass rate was a mere 39% (which is less than half the sub-test average), while the non-native pass rate was 70% (which is 10% higher than average in this group). Perusal of the answer records will surely provide us with an explanation of this rather curious close of the test.

The first clue is afforded by the fact that there was a profusion of variants of answers in the native group. No less than 79 different responses were elicited, while the non-native students delivered only half that number. (Ordinarily, it was the other way round.) This circumstance may probably be taken to indicate that the native students experienced problems of one sort or another (for instance a sudden loss of incentive to take heed of the prompts provided).

A further striking feature of the results was that the non-native group produced a very limited number of correct response types (the total number of different types of answers was relatively small, as indicated above). The vast majority (69%) of those who completed the question correctly used one and the same stock phrase: 'Do you mind if I'. In the native group, there was a much richer mixture of possible solutions (including 'Would it be asking too much to', 'You don't mind if I', 'Would it be inconvenient if I' etc). The most obvious way out, i.e. 'Do you mind if I', was chosen by 48 native students (corresponding to 31% of the total). All the students in the advanced group used this expression.

But there was also, as indicated, a very wide variety of impossible entries in the native group. A check of the answers showed that this was primarily due to the fact that students did not read the text after the gap carefully enough (and apparently not at all in some cases). Answers like 'Is it all right if I', 'I don't suppose I', 'Please may I' and 'Could I please' point in this direction. Some students did not even notice the question mark at the end of the line as evidenced

by attempts such as 'I will have to', 'So I had better', 'I think I'll' and 'I'll come back later if I'.

The risk of obtaining this negative effect is of course greater in cases where the gap is placed in the early half of the task. Placing the gap as late as possible is therefore always a worthwhile endeavour when writing test items of the present type.

To conclude: From the types of answers received we may infer that the abrupt end to the superior performance of the native students was probably caused by flagging motivation, manifested in lack of attention to all the attributes of the task. However, the way students reacted also highlights a disturbing weakness that may easily creep into this type of item. Anyone who does not, in item 14, register (consciously) one particular word, 'No', among close to 30 others may just as soon opt for 'Is it all right if I' (in itself a splendid way of asking permission) as 'Do you mind if I' - and thereby draw a blank! Swedes as well as Englishmen did so, the latter more often than the former. This seems to indicate that this task (alongside with item 4 discussed earlier) tends to reward a "premeditated" type of strategy in the use of the language; careful consideration and deliberation, rather than impressionistic reaction and spontaneity, appears to be the approach that is most likely to pay off. Looked at from a pragmatic and functional point of view this is not a very satisfactory condition.

9.3.11 Summary

The last few pages have been devoted to a scrutiny of various individual ways of responding to tasks in the Vocabulary-Grammar Test. It has been found that Swedish students' grammatical, lexical, and functional skills are highly variable, between students as well as across areas within the various skills. The area which caused most problems was idiomatic

usage. Here the native students were very much better, of course. Within the area of grammar, certain problems or "teaching points" turned out to be decidedly troublesome, for instance the 'it'/'there' distinction. Grammar was, however, less of a problem than idiomatic phrases.

As far as validity considerations are concerned, it was noted that the gap ought to come as late as possible in each item, that successful completion of the task ought not to hinge upon attention to little details in the text and, finally, that a slightly more elaborate marking scale would probably enhance the efficacy of the test substantially.

10 ATTITUDES

Attitudes were measured by means of two questionnaires directed to the English and Swedish teachers involved in the administration of the test. The English version (see Appendix 4) was answered by all those who were involved in the experiment in Manchester, i.e. 7 native English teachers. The Swedish questionnaire was completed by some 90 teachers of English from various parts of the country. The group represented approximately a tenth of the total number of teachers involved and did not constitute a random sample.

10.1 English teachers

In response to the item in which the English teachers were asked to state their opinions of the test, and whether they considered it a valid measure of English language skills, the following answers were received (one teacher did not answer this question):

> "Yes, but I think the essay section* actually reveals more about their mastery of the English language than do the other types of tests."
>
> "Yes, I would say that it is a valid test of foreign language skills."
>
> "It seems to demand a very idiomatic command of the language."
>
> "The test seems very well thought out and tests to a high level of ability: the comprehension exercise seems particularly exacting in the preciseness of thought and language it requires. Problems may arise with the Integrative Test because of idiomatic usages common to this area, and, possibly, age-group. The students do not always seem to be acquainted with the idioms that clearly the sentence required for completion."
>
> "More colloquial than expected with far greater use of idioms than in the teaching of foreign languages in England. This seems a sensible emphasis for the age-group serving as the target."

"Quite a difficult test with part 1 (the Reading Test) particularly taxing in that the wording of the answers requires logical thinking as well as precise understanding of the English. The Welsh accents will perhaps below the belt for Swedish students, etc.

(* I.e. the optional composition test, which the English teachers were informed about but never used in their groups.)

As the answers show, attitudes towards the test were quite favourable among the native teachers. They were impressed by the high level of proficiency which the test content reflected and noted in particular the use of difficult idiomatic language. Two teachers thought that the comprehention parts were particularly exacting and expressed some concern that the tasks require "preciseness of thought" and "logical thinking" as well as exact understanding of the language.

10.2 Swedish teachers

The opinions expressed by Swedish teachers were also mostly positive. There were quite a few comments about the level of the Vocabulary Test (1:1), which was considered too high by many. Some teachers were a little critical of the number of verb phrases ('make do', 'cut down', 'put up with' etc) included.

The Integrative Test (1:2) was generally very well received, although contracted forms (such as 'he's', 'I'd' etc) did not seem to be very popular. Some respondents thought they ought not to be approved of at all (in writing), others that their acceptance violates, logically, the one-word-per-gap rule.

Nothing much was said, really, about the first Reading Comprehension Test (2:1, the long text). Comments were succinct and mostly favourable (although not overly so). Lack of time was reported in some cases. There were hardly any complaints about the level of difficulty.

As regards the second part of the Reading Test (2:2), difficulty was, on the other hand, a major worry. Many teachers were of the opinion that the items tested at too high a level for the target group. Other than that, there were few negative comments.

The next sub-test, the Listening Comprehension Test (3:1), caused a whole host of distinctly negative reactions. The tenor of the message was that it is unfair and generally deplorable that regionally coloured English (an accent), enunciated at high speed and under emotion, should be used in a language test for schools. Individual comments ranged in quality from an unengaged "OK" to agitated outbursts such as "lousy" and "the qualifications of those who produced this year's LCT must be seriously called into question". According to some teachers, there were some adverse feelings among students, too. The majority of the Swedish teachers who sent in the Questionnaire therefore came to the conclusion that the listening comprehension task was very unfortunate this time and expressed the view that "dialects" should not be allowed in future tests.

The Vocabulary-Grammar Test (3:2), finally, went down quite well with the teachers, although there were several angry attacks on one particular item (No. 4 'there's no point in'; cf the analysis of results in Section 8.3). There were few comments on the level of difficulty, which may seem a little surprising in view of the fact that the sub-test did not belong to the easier ones.

A print-out of all the answers produced by the Swedish teachers (in Swedish) may be obtained free of charge from our Department.

11 SUMMARY AND DISCUSSION

11.1 Resumé of the experiment

The work described in the present report replicated an earlier study (Oscarson, 1986) which sought to determine the validity of the 1983 version of the National Test in English ('Centrala provet i engelska'), a general proficiency test used in the academically oriented Upper Secondary School in Sweden. The experiment was an attempt at construct validation of the test (the construct being the sort of English language ability which native speakers possess).

The primary aim of the present replication was to determine the construct validity of a later version of the same test (given in 1985). The method employed was a quantitative (statistical) analysis of the results obtained by a group of native English students who had been asked to take the test. The assumption behind the experiment was that educated native speakers would be able to reach very high scores on an English proficiency test which has claims to high validity. If the students were found to have difficulty in responding accurately to the test items, this would consequently be interpreted as a sign of poor construct validity.

A secondary aim was to study results on open-ended items in the test in order to assess some aspects of the written production skills acquired by a random sample of students for whom the test is designed. This qualitative linguistic analysis of answers was of interest mainly because the sample could be regarded as representative of the whole student population.

The native sample consisted of 166 English students at a Sixth Form College in Manchester. The average age of the students

was 17, and they were pursuing studies for O L[illegible] A Level examinations in various subjects. The sample represented a cross-section of the student population in respect of academic and linguistic abilities (i.e. students of below as well as above average ability were represented).

The Swedish group consisted of a 10% random sample of the total population of 34,000 students that took the test in 1985. The experimental group thus comprised 3,400 students. For the analysis of the productive skills a random sub-sample of 176 students was used.

The test consisted of sub-sections measuring - partly discretely and partly conjointly - vocabulary, grammar, reading comprehension, and listening comprehension, as well as language notions and functions. Both receptive and productive skills were assessed. The total number of items in the test was 100 (= total number of points awarded).

11.2 Main findings

The outcome of the research may be summarized as follows:

The native students achieved significantly higher scores on all parts of the test except one (a reading test). On average, they were correct on 83% of the test items. The average Swedish score was 61%. The high native score was taken to warrant the conclusion that the test is a valid measure of English language proficiency. The conclusion was reinforced by the outcome of a separate analysis of the results obtained by an "elite" group of native students.

A further major result was that the component parts of the test functioned quite differently with regard to discrimination between native and non-native ability. The Vocabulary Test (1:1) was the most sensitive of the six sub-tests, as the proportions of correct responses showed: the nati[illegible]s here

scored their highest average (92% correct), the non-natives their lowest (54% correct). The figures were about the same in the first study (carried out in York).

The most remarkable find appeared in the Reading section, in sub-test 2:1 (the long text). In conspicuous contrast to the situation in the Vocabulary Test, the English students here recorded their lowest average while the Swedish students recorded their highest. The result was that the two groups did not differ at all in terms of test scores. (There was a very small difference, but it was actually in the wrong direction.)

A surprisingly small difference was also obtained in the Listening Comprehension Test. The native students were only some 15% better than the non-native students. It was concluded that this is hardly a fair representation of the actual difference in ability to understand spoken English.

The remaining sub-tests displayed mutually similar (and average) result patterns.

As regards the secondary aim of the study (investigation of Swedish students' English language skills in absolute terms), it was found that there was very little correspondence between native and non-native performance across items. This may signify a difference in the structure of skills between the two groups. Certain problems, notably items involving idiomatic expressions, were disproportionately more difficult for non-natives than for natives.

Knowledge of English structures was rather uneven in the Swedish group, i.e. certain areas were mastered very well, whereas others (sometimes quite basic ones) were apparently not at all under control. On the whole, however, grammar was somewhat less of a problem than idiomatic phrases.

Attitudes towards the test were favourable among both English and Swedish teachers. English teachers were impressed by the

high level of proficiency which the test content reflected and noted that the use of idiomatic language was not shied away from. The comprehension parts were judged to be particularly exacting and some concern was expressed that the tasks require "preciseness of thought" and "logical thinking" as well as exact understanding of the English language.

Swedish teachers, too, were in the main pleased with the test, and the critical comments were mostly on details. There was one very serious objection, however, and this concerned the Listening Comprehension Test, which was considered unsuitable by a majority of respondents. Non-standard pronunciation and noisy acting were the main complaints lodged.

11.3 Discussion and conclusions

In this part of the report, we try to piece together and discuss some of the main strands of our research. We will not repeat figures and previous discussion, but frequent reference will be made to relevant sections and tables in the foregoing chapters.

11.3.1 The native score level

Recurning first to the test scores (cf Table 2), let us consider the overall native performance level for a moment. It might have been expected, perhaps, that the English students would have scored much closer to the 100% correct response ratio than they did and that the difference in relation to the Swedish students would thereby have been larger than it turned out to be. While this is a highly natural and plausible hypothesis, we must recognise that there are at least two factors that tend to work against the perfect average score, no matter how proficient the test-takers may be: chance variation (due to the occasional lack of attention, for instance) and less than total understanding of - and familiarity with -

testing procedures and instructions (and the intentions behind them).

Chance variation (due to faltering attention) may in the case of native speakers result, quite simply, from boredom, because the task is often not demanding enough. It may also, as with any other group of testees, result from external interference, distracting noise and suchlike (a case in point being a disturbance caused by "tree felling with a chain saw ... near the classroom" which was reported by one of the staff in Manchester on one occasion). In brief, as Stern (1983) notes, "while all native speakers possess communicative competence in their first language ... they will at time use the language inappropriately and commit 'faux pas' or 'drop bricks' (p.345)". Therefore one should always, on this count alone, take a little percentage off the theoretically expected score in order to arrive at the more realistic level which an obviously over-qualified audience is likely to reach.

Furthermore, in an experiment like the present one, some allowance must be made for the usually less than maximal opportunities that native subjects are offered for practice on the particular types of task at hand. Our Manchester students were as well prepared for their job as one could reasonably expect, but if they had had the same amount of previous experience with relevant materials and routines as their Swedish counterparts they would undoubtably have advanced several rungs on the 100-point ladder. (Most Swedish students will have been given one or more trial runs with previous tests before they sit for the real thing.)

Lastly, there is also the question of the natural variation of language proficiency, i.e. even in native samples. Not all natives are able to use their language flawlessly. In view of the level of the test, it may be assumed that some of the tasks were genuinely difficult for some of the English students.

What the above discussion amounts to is the following. We never expected the British youngsters to perform at the 100% level on our English test. Our considered estimate was set some 10% lower. However, as this figure only goes some way, but not all the way, towards equating the expected performance level and the level actually attained (83%), we have reason to believe that there is still some scope for improvement as far as test validity is concerned. Particularly in the area of reading comprehension, this would seem to be a plausible assumption.

11.3.2 Reading comprehension

Reading comprehension is measured, directly, in sub-tests 2:1 (the long text) and 2:2 (ten mini-texts), and also, somewhat more indirectly, in sub-test 1:2 (the Integrative Test). Furthermore, although this is not explicitly stated or intended, reading skills come into play in sub-test 1:1 (the Vocabulary Test) and 3:2 (the Vocabulary-Grammar Test). Even in sub-test 3:1 (the Listening Comprehension Test) a modicum of reading comprehension is required in that the response options in the test booklet must be read and understood before correct answers can be delivered. Thus the ability to read and understand the language is a most essential prerequisite for successful performance in the test, and this is not at all uncommon in a test of the kind we are dealing with here. Actually, it would be very difficult to manage the testing task at hand (which is quite formidable) without making extensive use of textual material. Having said that, we might add that reduction of the degree to which facility with written discourse determines test outcomes would still seem to be a worthwhile goal to pursue, not least in view of the importance now attributed to aural-oral skills (and in view of the fact that the National Test sets a standard which has considerable influence, for better or for worse, on the language teaching scene).

Another aspect of test content that has to be considered in this context is that of sampling. It is of course imperative that texts included are unequivocal reflections of reading matter envisaged in the Curriculum, and there can be no doubt that this is the case in the National Tests. All samples used are very safely inside the boundaries of curricular specifications; only texts within a relatively limited and fairly well-defined range of written discourse (typically, straightforward non-specialized narrative fiction and prose) are used and this is of course in principle a very good thing. Teachers and students alike can always rest assured that there will be no surprise shocks in store for them in the way of unexpected types of text and they can confidently prepare themselves for any upcoming round of national assessments. All this is entirely fair and unobjectionable, and the system makes for smooth co-operation between the parties involved in the undertaking, and thereby for efficient execution of a difficult task.

The other side of the coin is that there is a great deal more to reading comprehension than just the ability to comprehend passages of narrative prose of a general and predictable kind. That is, the construct of reading comprehension (cf Chapter 2), as conceived of in our study, and probably as understood by the general public, relates to the ability to interpret written language in a wider sense, i.e. irrespective of level, genre, style, topic, register etc. By comparison, the goal of reading comprehension in the Curriculum, emphasizing understanding and appreciation of literature and ordinary prose, is actually rather limited. This circumstance has consequences which should be borne in mind when the capabilities of our native and non-native samples are being compared (and when the validity of the Reading Test is being considered). Equality of scores, which did occur in one case (cf Section 8.2), can hardly be taken as proof of comparable overall reading skills, precisely because the tests do not measure reading comprehension globally. The question of whether equal scores should be taken as counter-evidence of test validity will be given some attention below.

The test which resulted in equal scores was the long text followed by comprehension questions (2:1). This sub-test was the hardest of all for the native students (while at the same time it was the easiest for the non-natives, cf Figure 2). Not even the most advanced native group, who had actually been given extra incentive to do their very best (other students having failed to perform up to expectations), managed to demonstrate convincing ability (cf Table 3). As will be recalled (cf Table 1), the York students also found the long text (a different one) troublesome. Obviously, English language comprehension in itself, at least not ordinary decoding skills, which our native subjects unquestionably possess in ample measure, will not suffice as a basis for excellent performance on this test. What else, then, may be needed, and to what extent can the test be regarded as a valid measure of reading comprehension?

The intention behind the comprehension questions is to gauge overall understanding of text meanings (referred to by Widdowson, 1983, as "indexical meaning"), while avoiding tasks which require only superficial semantic deciphering of individual words, phrases, and sentences (i.e. "symbolic meaning" in Widdowson's terminology). This is in line with statements in the Curriculum to the effect that, at the present stage of language learning, concentration on attention to form in the study of texts should gradually give way to more emphasis on appreciation of content. Logic requires, then, that questions should be designed in such a way that their solution can only be arrived at through a process of perceiving and amalgamating sets of contextual clues, rather than comprehending isolated items of information. If we are successful in achieving this aim, it follows that we are moving into an area where non-language-specific variables such as deductive ability, background knowledge related to the topic (or knowledge of the world), associative memory, reasoning etc become increasingly important and where we, therefore, should expect a smaller difference between native and non-native test scores. Indeed, this is what happened in our experiments.

However, as long as the English language is the medium of the message, and as long as language-independent factors can only explain part of, and never all of, the variance on our reading test, we should hardly expect English and Swedish students to perform on a par with each other. We must conclude either that the English sample is motivationally or intellectually inferior to the Swedish sample or else that the test is not as valid a measure of reading comprehension as it might perhaps have been. In view of the fact that not even the highly intellectual portion of the English sample (cf Section 8.3) managed to reach a very high correct score level, and in view of the fact that the York and Manchester studies converged at the very modest 70-80% level, the latter conclusion seems to be more plausible than the former.

If the above assumption is correct, the next question to consider is this: What can be done in order to improve the validity of the reading test? First of all, it must be emphasized that the overriding goal aimed at, that of grasping the overall meaning of pieces of written discourse, or comprehension at a deep level, cannot and should not be called into question. It represents ultimate skills of great importance. On the other hand, it would seem that the language component (to the extent that it may be separated from the generalized types of abilities referred to earlier) ought to be allowed to play a more significant role. That is, if modifications of the test type were to be contemplated, they ought to go in the direction of linguistically more demanding texts. Balancing this measure, while still striving to emphasize sensitivity to "pure" comprehension, one ought to simplify the question apparatus, e.g. by distinguishing more clearly between given multiple choice options (if such a task format is used) thereby avoiding distractors which are dangerously close to a correct answer. As it is now, the best test-taking strategy may very well be to read the question and options first and then, by a process of meticulous comparison and matching, find the answer in the relevant paragraph. It hardly needs to be said

that this type of behaviour has very little to do with reading for overall understanding or for, say, literary appreciation.

The simple logic of the point made is that a test of reading comprehension should yield the suitable spread of results not on account of the fact that the alternative answers to multiple choice questions are semantically or conceptually difficult to choose between, but rather on account of the fact that the text (the language) is difficult to understand. The questions should in fact be worded in relatively simple terms, and they should in any case be more easily mastered by testees who are, overall, more proficient than those for whom the test is designed.

To end this discussion of the first part of the reading comprehension section, we will venture the prediction that the validity of the test would increase if a greater diversity of text types were employed (within the confines of curricular recommendations, of course). The long text (of approximately three pages) might for example be replaced by two shorter ones representing different genres or topics. In all probability, such a measure would provide a better basis for reliable and valid assessment of the skill in question.

The second part of the reading comprehension section (sub-test 2:2) measures reading more directly, while at the same time rather more superficially, than the first part. Judging by the performance of the native speakers, as well as by the statistics (cf Table 2), the test is a valid one. It is also, one might say, "cost-effective" in that it is less time-consuming than most other types of reading comprehension tests (e.g. the type discussed above). The correlation with the first part (2:1) is not particularly high (r = .59 in the Swedish group, cf Appendix 6), which indicates that the two tests partly measure different aspects of the tested skill. Taken together, these facts provide strong support for retaining, and possibly expanding, sub-test 2:2.

11.3.3 Listening comprehension

The testing technique used in the Listening Comprehension Test is basically the same as in the long Reading Comprehension Test, i.e. it involves multiple choice questions on the content of a piece of discourse (spoken discourse, naturally, in the case of the former). Any weakness spotted in either test is therefore likely to show up, at least occasionally, in the other, and, as we have seen, the non-difference obtained on the Reading Test in Manchester has its analogue in the results obtained on the Listening Test in York (cf Section 3.1). We assume, therefore, that the conclusions drawn above concerning the Reading Test are, in certain respects, applicable to the Listening Test as well. This means, for example, that the relatively small disparity observed in test scores between the two groups of students is judged to be disproportionate to the actual difference in ability to understand the language. In reality, the natives and non-natives doubtlessly differ to a much larger extent, the reason being that the Listening Test only measures - and this is hardly a revelation to anyone concerned - comprehension within quite narrow bounds of speech realization (normally RP English in a generalized narrative mode) and within which the Swedish students have had most of, in some cases all of, their aural training. We must recognize, therefore, that the pleasing picture of the Swedish students' ability to understand spoken English, in 1983 (York study) as well as in 1985 (Manchester study), is at least partly an effect of artificially "inflated" test results.

The resemblance between the reading and listening tests, with regard to structure as well as outcome, would seem to justify the further parallel conclusion that rather more variation in input (i.e. in respect of types of recordings used) would be beneficial to test validity. Thus two separate sets of tasks, instead of a single unitary set, representing for example British and American English, or formal and informal English, or dialogue and descriptive (or narrative) exposition, or some other such pair of complementary linguistic representations,

might be used in order to ensure more valid listening test results.

A further question worth considering, in view of the great importance attached in the Curriculum (1970, II. p 265) to practical language skills, is that of a possible expansion of the number of tasks measuring listening comprehension. At present, listening accounts for a little more than a tenth of the total number of points available, while reading, writing, knowledge of words and phrases, and related skills, take up all of the remaining points. Increasing the weight of the listening score does seem justified in this perspective. We would suggest, furthermore, if such a step were to be taken, that listening tasks of a mini-context type be used, i.e. tasks analogous to the ones used in the second part of the Reading Test (sub-test 2:2, cf Section 5.2). These would then measure understanding of restricted utterances, or spontaneous and immediate understanding, and would serve as a natural supplement to the more searching and global type of questions asked in the current test.

Finally, we will return for a moment to the matter of suitable speech styles in a listening test at this level. As was noted in Sections 8.2 and 10.2, the Swedish teachers came down very heavily on the present test, essentially because the language used was tinged with a Welsh accent.

It is interesting, however, to look at the results and to compare the record of the present test with that of the 1983 version of the test (cf Section 2.1), which constituted a straightforward representation of "received pronunciation" (RP) delivered at a pedagogically suitable rate of speech. In 1983, the Swedish LCT score level was on a par with, or slightly above, the total average level. In 1985, the LCT level was decidedly higher than the average level (which in both years corresponded to 61% of the maximum score). That is, Swedish students did in fact do better on the more authentic (and much criticized) version of the test. Not surprisingly,

this was also true in the case of native speakers. In 1983, the English group achieved a listening comprehension score which was way below their total average. In 1985, the native LCT score was on the same level as, or even above, the total average. Reliability figures were also higher in 1985 (KR20 = .53, in the Swedish group, as against .45 in 1983), which means that the 1985 version of the test yields more stable (less inconsistent) results. On the other hand, the standard deviation was larger, i.e. better, in the 1983 test but only marginally so.

Thus our research evidence speaks in favour of the more realistic type of listening comprehension materials that the 1985 version of the test exemplifies. Nonetheless we must of course take very careful note of the sentiments voiced by practising teachers. After all, validity is but one important consideration when deciding on test content and format. Practicality, feasibility, and suitability are others. Therefore, if rather more authentic recordings were to be reintroduced (the likelihood of this happening is not very strong at the moment), better ways of presenting them would certainly have to be worked out. Allowing time for warm-up at the beginning of the tape, say 5 minutes, so as to give students a chance of getting used to, or tuned in to, voices, rate of speech, topic etc, would seem to be a very important first step. Further experimentation (not as part of the yearly national assessments, of course) would be another vital measure. Careful information about facts and figures, as well as explanation of rationale and objectives, would also be required.

11.4 Recapitulation of some key points

Below are recapitulated very briefly some of the key points in this report. References are to previous Sections providing more thorough treatment of each issue.

1. In an earlier validation study, carried out in 1983, native English students obtained high scores on the National Test in English. This is a sign of test validity. (3.1)

2. The English students were most successful on the Vocabulary Test, and least successful on the comprehension parts (Reading and Listening). (3.1)

3. Swedish students' formal command of English was very variable. Elementary vocabulary and grammar mistakes were not uncommon. The students' functional command of the language, as shown in the comprehension sections, was comparatively strong (3.1).

4. Further investigation of the test, and of the proficiency of Swedish students, was judged to be needed. (3.1; 3.4)

5. Similar investigations of French and German tests were undertaken in 1985. Both native French and native German students achieved very high scores. The results testify to the validity of the two tests. (3.2-4)

6. The native French students reached their highest scores on a dictation, and on tasks measuring grammar, vocabulary, and phrases. Open-ended tasks in the latter areas, as well as reading comprehension tasks, resulted in relatively low scores. (3.2)

7. The native German students obtained their best results on a test measuring grammar, vocabulary, and phrases, and on a test

of listening comprehension. Their reading comprehension score was comparatively low. (3.3)

8. For control purposes, the English validation was repeated in 1985, using a different version of the test and new groups of native English students (4-7). Again the native students obtained high scores (which confirms that the test is valid), and again their best result was on the Vocabulary test, whereas they did no better than the Swedish students on the Reading test which involved passage comprehension. (8.2)

9. In contrast, Swedish students achieved their highest score on the Reading test, and their lowest score on the Vocabulary test (8.2). English and Swedish students' average scores on individual items did not correlate well (9.2.8; 9.3.2). The results suggest that there are significant structural differences between the language skills of English and Swedish students.

10. In the "productive" sections of the test (sentence completion), Swedish students had most problems with idioms and certain points of grammar. (9.2.8; 9.3.11)

11. Both English and Swedish teachers liked the test. However, Swedish teachers criticized the Listening test. (10.1-2)

12. Although the test was found to yield valid scores, the outcome of the study suggests that there is still room for improvements. (11.3)

13. There is a risk that the comprehension sub-tests measure too narrowly in one respect (that of language represented) and too widely in another (that of abilities required for completion of tasks). (11.3.2-3)

14. The validity of the Reading test might increase if a greater variety of texts was used as a basis for tasks, and if the linguistic level of the textual material was raised. Texts

ought to be relatively difficult, questions relatively simple. (11.3.2)

15. Likewise, the validity of the Listening test might increase if more variation in respect of types of recordings was introduced. Increasing the number of listening comprehension tasks, as well as advancing authenticity, seems justified. (11.3.3)

REFERENCES

af Ekenstam, N.-H. (1986) Tyska elever och svenska tysk-prov: Hur klarar tyska gymnasister våra centrala prov i tyska? /"German students and Swedish tests of German: How do German Upper Secondary School students perform on our national tests in German?"/ 'Rapport' No. 1986:06, Department of Education, Gothenburg University, Sweden

Angelis, P. (1977) Language Testing and Intelligence Testing: Friends or Foes? Occasional Papers in Linguistics, No. 1, Southern Illinois University, Carbondale

Carroll, J.B. (1973) Foreign language testing: Will the persistent problems persist? I O'Brien, M.C. (ed), Testing in Second Language Teaching: New Dimensions ATESOL, Dublin: Dublin University Press, pp 6-17

Chambers, J.K. and P. Trudgill (1980) Dialectology Cambridge: Cambridge University Press

Cronbach, L.J. (1971) "Test Validation" In R.L. Thorndike (ed), Educational Measurement Washington: American Council on Education, pp 443-507

Crystal, D. (1985) A Dictionary of Lingusitics and Phonetics Oxford: Basil Blackwell

Curriculum (1970) Läroplan för gymnasieskolan: I. Allmän del. II. Supplement (The Upper Secondary School Curriculum) Stockholm: Liber Utbildningsförlaget

Davies, A. (1985) "John Oller and the restoration of the test" System, Vol. 13, No. 2, pp 99-104

van Ek, J.A. and L.G. Alexander (1980) Threshold Level English Oxford: Pergamon Press

Ferguson, G.A. (1966) Statistical Analysis in Psychology and Education London: McGraw-Hill

Guilford, J.P. (1965) Fundamental Statistics in Psychology and Education New York: McGraw-Hill.

Hellekant, J. (1986) Franska elever gör ett svenskt fransk-prov: Ett försök till validering av det centrala provet i franska 1985 /"French students take a Swedish test in French: An attempt at validating the 1985 version of the national test in French"/ 'Rapport' No. 1986:05, Department of Education, Gothenburg University, Sweden

Irujo, S. (1986) "Don't Put Your Leg in Your Mouth: Transfer in the Acquisition of Idioms in a Second Language". TESOL Quarterly, Vol. 20, No. 2

de Jong, H.A.L. (1983) Focusing in on a Latent Trait: An Attempt at Construct Validation by Means of the Rasch Model In J. van Weeren (ed), 'Practice and Problems in Language Testing 5', Non-classical test-theory final examinations in secondary schools (pp 11-35) Centraal Instituut voor Toetsontwikkeling (CITO), Arnhem, Holland

Klein-Braley, C. (1985) "A cloze-up on the C-test: A study in the construct validation of authentic tests" In Language Testing, Vol. 2, No. 1, pp 76-104

Lado, R. (1961) Language Testing: The Construction and Use of foreign Language Tests London: Longman

Lado, R. (forthcoming) "Analysis of Native Speaker Performance on a Cloze Test" Paper read at the LT +25 Language Testing Symposium in Honor of John B. Carroll & Robert

Lado, Quiyat Anavim, Israel, May 1986 (to be published in Language Testing, 3, 2, 1986)

Löfgren, H. (1969) Mätningar av språkfärdighet i tyska: En undersökning på elever i årskurs 7 /"Measuring proficiency in German: An investigation with pupils in grade 7"/ Pedagogisk-psykologiska institutionen, Lärarhögskolan i Malmö, Sweden

Oller, J.W. Jr. (1979) Language Tests at School: A Pragmatic Approach

Orpet, B.R. (1985) "Foreign Language Testing in Sweden" British Journal of Language Teaching, Vol. 23, No. 1, pp 37-41

Oscarson, M. (1986) Engelska och svenska elevers prestationer på ett centralt prov i engelska: En valideringS- och utvärderingsstudie. /"The Performance of English and Swedish Students on a Standardized Test in English: A Validation Study"/ Rapport nr 1986:02, Department of Education, Gothenburg University, Sweden

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik (1985) Comprehensive Grammar of the English Language London: Longman

Sang, F., B. Schmitz, H.J. Vollner, J. Baumert and P.M. Roeder (1986) "Models of second language competence: a structural equation approach" Language Testing, Vol. 3. No. 1 pp. 54-79

Stern, H.H. (1983) Fundamental Concepts of Language Teaching Oxford: Oxford University Press

Taylor, W.L. (1953) "Cloze procedure: a new tool for measuring reliability" "Journalism Quarterly", Vol. 30, pp 415-433

Svartvik, J. et al (1982) Survey of Spoken English: Reports on Research 1975-81 Lund: Gleerup/Liber

Svartvik, J. and O. Sager (1983) Engelsk universitetsgrammatik Stockholm: Esselte Studium

Widdowson, H. (1983) Learning Purpose and Language Use Oxford: Oxford university Press

A P P E N D I X 1

Information and instructions

Manchester Groups

GÖTEBORGS UNIVERSITET
Institutionen för pedagogik

GOTHENBURG UNIVERSITY
Department of Education and
Educational Research
Language Teaching
Research Unit

Validation of tests in English
Manchester 1985

INFORMATION AND INSTRUCTIONS

Dear Colleague,

First of all we would like to thank you for your assistance in this validation study. The testing will be of very great value to us in our attempts to improve the national language tests used in Swedish schools.

AIM

The main aim of the assessment in Manchester is to determine the average performance level reached by native speakers in each of the sub-tests in our national test in English. This will help us establish the validity of our present testing procedures.

INFORMATION TO STUDENTS

We would be grateful if teachers would inform the students about the purpose of the testing. The outcome will help us develop our national language tests in the right direction. Basically we want to compare the results obtained by native speakers and the results obtained by our own target group, i.e. learners of English as a foreign language in the upper secondary school ("sixth-formers").

We will be happy to send you individual results, as soon as we have done the marking, if the students are interested. We will also be pleased to answer any further questions about the assessment under the address above.

Finally we would appreciate it very much if you would convey our thanks the students for their willingness to take part in this research.

stadress
x 1010
31 26 MÖLNDAL, Sweden

Besöksadress
Frölundagatan 118
Mölndal

Telefon
Nat 031-67 90 00 växel 031-67 ... direktval
Int +46 31 67 90 00

THE TEST

The test is the 1985 version of the National Test in English which is taken by all students in the upper secondary school at the age of 17. The function of the test (when used in Sweden) is to ensure comparability in marks awarded in different schools throughout the country.

The structure of the test is as follows:

Sub-test	Time allowed
TEST PAPER 1 1 Vocabulary Test 2 Integrative Test	35 min
TEST PAPER 2 Reading Comprehension Test, Parts 1 & 2	35 min
B r e a k The students leave the room.	15 min
TEST PAPER 3 1 Listening Comprehension Test 2 Vocabulary-Grammar Test	35 min

(The test also contains an essay part but this is not included in the validation process in Manchester.)

The times given are those which Swedish students are allowed for the completion of each sub-test. English students will of course be able to complete the tests, with the same degree of concentration, in less time than our own students in Sweden.

The following materials are provided:

"INFORMATION AND INSTRUCTIONS"
TEST PAPER 1
TEST PAPER 2
TEST PAPER 3
SOUND TAPE (Open reel or cassette)
QUESTIONNAIRE (for teachers/invigilators)

TEST PAPER 1

1 The test booklets are placed on the desks before the students are allowed into the room.

2 The students are asked to fill in their names etc on the front page of th. booklet. (We need their names in order to be able to calculate individual aggregates.) Students should not open their booklets while instructions are being given.

3 The teacher then gives the following information:

The instructions for this test ar· i. thi· booklet. The answers to the first tasks (Vocabulary Test) are to be given in the numbered boxes at the bottom of each page. The second part of the TEST PAPER (Integrative Test) consists of a text in which certain words have been deleted and replaced with blanks. Your task is to insert the words that have been deleted.

4 When the students seem to be ready, or time is up, the teacher hands out TEST PAPER 2.

TEST PAPER 2

1 The teacher informs the students:

The test is in two parts. The instructions are in the booklet (TEST PAPER 2). Write your answers in the first booklet (TEST PAPER 1, page 11).

2 After TEST PAPER 2 there is a break. The students leave the room.

TEST PAPER 3

1 It is very important that the listening comprehension test be administered under favourable listening conditions. A good tape recorder is needed and it should be checked beforehand. The room must be suitable from an acoustic point of view and it should, ideally, be of ordinary classroom size. Testing in a large room, e.g. a lecture hall, is not recommended.

2 The booklets (TEST PAPER 3) are distributed before the students return.

3 The students are asked to fill in their names etc on the front page. They are also informed that the tasks in the listening comprehension test are multiple-choice and that they will be given time to transfer their choices to the boxes on page 3 after the tape has been played.

4 Start the tape recorder and listen to the first sentence. Adjust the volume. Rewind the tape. Tell the students that the instructions are on the tape. Start the tape recorder again.

In case of unexpected disturbance the tape recorder may be stopped. The tape is then rewound slightly and started again.

5 When the Listening Comprehension test is over the teacher reminds the students that they are supposed to transfer their answers to the boxes on p.3. They are then asked to do part 2 (Vocabulary-Grammar Test) on p 4.

6 Finally all test papers are collected.

ONCE AGAIN OUR SINCERE THANKS TO YOU AND YOUR STUDENTS FOR YOUR KIND COOPERATION IN THIS VALIDATION STUDY.

Dr Mats Oscarson
Coordinator of the study

A P P E N D I X 2

The Integrative Test (Subtest 1:2)

THE NATIONAL BOARD
OF EDUCATION

UNIVERSITY
OF GOTHENBURG
SWEDEN

NATIONAL TEST

IN

ENGLISH 1

FOR THE UPPER SECONDARY SCHOOL, 1985

TEST PAPER 1

AND ANSWER SHEET FOR SUB-TEST 2
READING COMPREHENSION TEST, (P. 11)

SUB-TEST 1: INTEGRATIVE TEST

NAME: _____

SCHOOL: _____

COURSE OF STUDIES: _____

PART TWO: Integrative Test

Instructions

1. *Study the text and fill each blank with ONE word.*
2. *Any contracted form, such as can't or it's, counts as ONE word.*
3. *Try to fill in all the blanks.*

Kenneth and Pamela have known each other for a short time. They have been to the cinema together once or twice. The last time they had a date Pam was unable to keep it and Ken waited for her in vain. As soon as she could she phoned him and apologized, explaining why she hadn't turned up. Now they have met again.

KEN: Tell me a little about your family, Pam. For instance, what _____(1)_____ your dad do?

PAM: He's an engineer. His job takes him all over the country, and abroad too, sometimes, so he's hardly _____(2)_____ at home. Mother says it's like _____(3)_____ married to a sailor.

KEN: Yes, I can imagine...

PAM: Then again she says it _____(4)_____ its advantages. You never have a chance to get fed _____(5)_____ with a husband who's only at home occasionally. They're like a couple of proper love-birds when he _____(6)_____ turn up. You'd think _____(7)_____ been married only a month instead of twenty years...

1202-01

Of course _____ (8) be different when she hasn't got me for company. Be a bit lonely for her then.

KEN: Are you thinking of _____ (9) home, then?

PAM: Well, I suppose I shall one _____ (10), when I get married, I mean.

KEN: How old are you, Pam?

PAM I was seventeen last Christmas.

KEN: You're only a kid, Pam. You _____ (11) be leaving your mother for a while, will you?

PAM: Well, a girl's got to think about the future, _____ (12) she? Many a girl's got married and started a family at eighteen. Anyway, how old are you, Old Greybeard, if it isn't _____ (13) personal a question?

KEN: Twenty. And what sort of chap are you going to marry? Somebody like your dad _____ (14) away most of the time?

PAM: No fear. I'll want my husband to be with me all the time and I'll risk getting _____ (15) of him.

KEN: Don't be too sure. You'd better wait _____ (16) he turns up. He might _____ (17) out to be a sailor or something.

1202-01

PAM: How do you know he _____ turned up already?
18

KEN: Oh... (*Pause*) Well, what are you doing out with me, then?

PAM: I went out with you just to make him jealous.

KEN: I see. Now this future husband of _____, is he a
19
great big bloke?

PAM: Oh, I wouldn't say that. He's quite well-built, though.

KEN: Good _____ fighting, is he?
20

PAM: I should think he can take care of himself.

KEN: Hmmm. (*Pause*) Well, good night then.

PAM: (*Laughs*) Come on, Ken. I was only _____ your leg.
21

KEN: Oh, I knew that all the time, of course. I only pretended to be fooled.

PAM: Smart, _____ you?
22

KEN: Immensely.

(*A short silence*)

PAM: When I didn't turn up last night, did it occur _____
23
you that I might have got held up somewhere?

KEN: It did cross my _____.
24

PAM: You didn't think I'd made the date and then deliberately not turned
up, _____ you?
25

1202-01

KEN: It has been known _____ (26) happen, you know.

PAM: Well, you don't know me very well if you think I could do a thing like that.

KEN: Well, it's not _____ (27) if we were old pals, it it? And when you turned up with that Christine the other night...

PAM: I certainly didn't want her to come, you know. Only I couldn't get _____ (28) of her without offending her. Christine is _____ (29) that, you know. She got it into her _____ (30) that she was coming to have a look at you. She said _____ (31) only stay with us for five minutes and then go. And you know what happened.

KEN: Look, Pam, I didn't mean to tear her _____ (32) pieces like that, you know; only all _____ (33) insinuations of hers made me furious. I just couldn't _____ (34) telling her exactly what I _____ (35) of her. So when all that happened and you didn't turn up last night, well, I just thought you didn't want to see me any more and you didn't like telling me to my face.

1202-01

PAM: And it wasn't that way at all! Doesn't it just show how misunderstandings can come about?

KEN: Well, it's all history by now. Let's go and have some coffee, shall we?

If there is time left, go back and check your answers.

A P P E N D I X 3

Vocabulary-Grammar Test (Subtest 3:2)

PART TWO: Vocabulary-Grammar Test

Instructions

1. *In each of the following 14 mini-texts there is a blank indicating that two or more words are missing.*

2. *Study each text, and then put in the missing words so that it makes good sense and is correct English.*

3. *As a rule, 2 - 4 words are enough to complete the sentence. There should not be more than six.*

- -

1. JIM: This advertisement says that the machine is "fool-proof". What _____ by that, Daddy?

   DAD: That it's so simple that anybody can handle it, even a fool.

2. As soon as I saw the new manager I thought there was something familiar about him. I knew _____ before, but I just couldn't remember where.

3. It's quite clear that Tom messed up the deal, but he's learnt his lesson by now. I'm sure _____ same mistake again.

1202-03

4. LEN: The damage is done
and _____
in worrying about the consequences now.
RON: That's easy for you to say.

5. If Jimmy had a lot of money, I'm sure
_____ himself
a veteran car.

6. ANN: Have you asked your parents if you can go
mountain-climbing with me in Norway?
PAT: Yes, and I'm afraid
they _____ to go,
because they think it's too dangerous.

7. DAVE: It's irritating that the boss refuses to discuss
our project.
ALAN: Yes, isn't it? I've tried to
make _____ mind
a couple of times, but he won't.

8. TONY: I'm awfully tied up at the moment, so I
can't help you.
TED: Why didn't you say so yesterday when I asked you?
If you had told me you were so busy,
I _____ else.

1202-03

9. TONY: Who looks after Marilyn when you're away?

MAUD: A friend of ours.

TONY: You don't use that baby-sitting agency?

MAUD: No, Marilyn hates _____ by someone she doesn't know.

10. ROBIN: The forecast says that _____ all over the country tomorrow and probably for the rest of the week.

OSCAR: Oh dear, more rain! I was hoping for some sunshine for a change.

11. LINDA: What a victory! We ought to celebrate.

BRIAN: Yes, _____ restaurant for a really good meal.

LINDA: That's an excellent idea.

12. HELEN: Simon is good at German.

DIANA: _____ fluently?

HELEN: Oh, yes, you'd think he was a native.

13. JOAN: Have you seen a film called "Total Eclipse"?

TESS: Yes, unfortunately. It's _____ seen. I've never been so bored in a cinema.

14. SECRETARY: I've got to pick up my child at the nursery today. _____ leave early?

MANAGER: No, that's all right. Have a nice weekend.

SECRETARY: Thanks. You too.

If there is time left, go back and check your answers.

1202-03

A P P E N D I X 4

Questionnaire, English Teachers

GÖTEBORGS UNIVERSITET
Institutionen för pedagogik

GOTHENBURG UNIVERSITY
Department of Education and
Educational Research

Validation of Tests in English
Manchester 1985

Q U E S T I O N N A I R E

We would appreciate it very much if teachers participating in the validation experiment would answer the following questions after the administration of the test.

1 Teacher and/or group: .....

2 How long did it take the students to complete the test?

.....

3 Was there any kind of disturbance (or any other problem) that may have affected the students' performance?

.....

.....

.....

4 How would you characterise the group(s) in respect of academic and/or linguistic ability?

.....

.....

.....

5 What is your opinion of the test itself (bearing in mind that its chief function is to assess group means)? Would you say that it is a valid measure of foreign language skills?

.....

.....

.....

.....

.....

THANK YOU VERY MUCH INDEED FOR YOUR VALUABLE HELP!

Postadress
Box 1010
S-431 26 MÖLNDAL, Sweden

Besöksadress
Frölundagatan 118
Mölndal

139

Telefon
Nat 031-67 90 00 växel 031-67 ..... direktval
Int +46 31 67 90 00

APPENDIX 5

Frequency Distribution of Test Scores, Swedish Sample

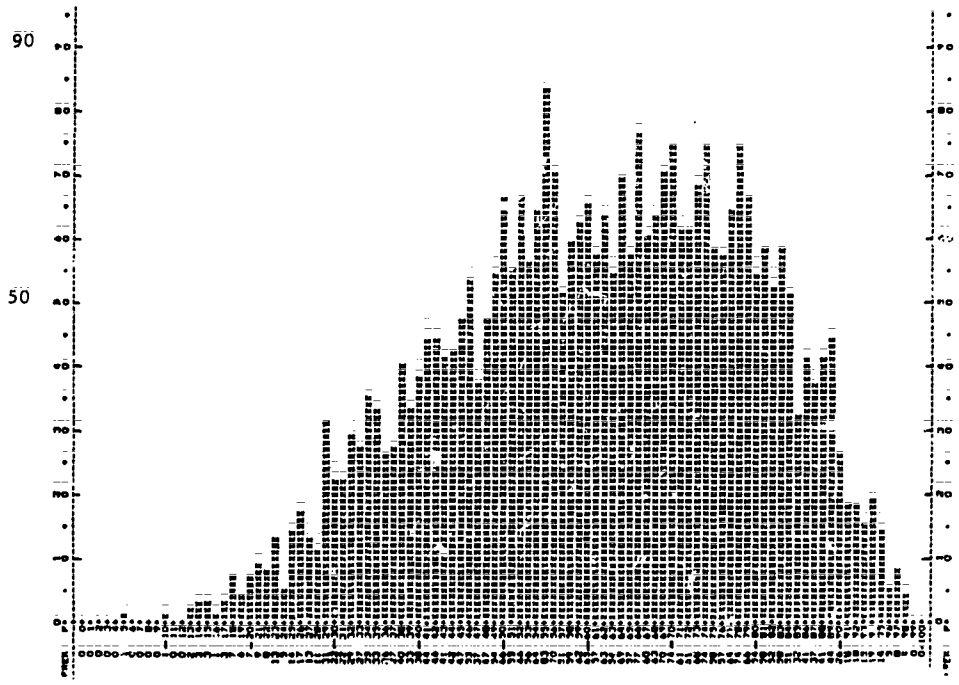


Figure 7 Frequency Distribution of Individual Test Scores in the Swedish Sample (N = 3,409)

A P P E N D I X 6

Intercorrelations among Subtests
English and Swedish Samples

Table 7 Intercorrelations among Subtests and Total Score: The Native Sample (N = 147)

	1:1	1:2	2:1	2:2	3:1	3:2	Total
1:1 Vocabulary (18)	1.0	.44	.39	.36	.38	.38	.66
1:2 Integrative (35)		1.0	.35	.35	.34	.57	.86
2:1 Read.Compr.1 (12)			1.0	.47	.34	.31	.65
2:2 Read.Compr.2 (10)				1.0	.31	.32	.62
3:1 List.Compr. (11)					1.0	.17	.54
3:2 Voc.-Gramm. (14)						1.0	.70
Total (100)							1.0

Table 8 Intercorrelations among Subtests and Total Score: The Non-Native Sample (N = 3,409)

	1:1	1:2	2:1	2:2	3:1	3:2	Total
1:1 Vocabulary (18)	1.0	.75	.60	.69	.50	.67	.87
1:2 Integrative (35)		1.0	.64	.67	.47	.79	.94
2:1 Read.Compr.1 (12)			1.0	.59	.44	.56	.75
2:2 Read.Compr.2 (10)				1.0	.48	.59	.80
3:1 List.Compr. (11)					1.0	.40	.60
3:2 Voc.-Gramm. (14)						1.0	.85
Total (100)							1.0

REPORTS FROM DEPARTMENT OF EDUCATION AND EDUCATIONAL RESEARCH, GOTHENBURG UNIVERSITY

issn 0282-2156

Available from: Dept of Education and Educational Research, Gothenburg University, Box 1010, S-431 26 Mölndal, Sweden

---

Reuterberg, S.-E. On comparing transition rate gains. 1985:01

Emanuelsson, I. & Svensson, A. Does the level of intelligence decrease? A comparison between thirteen-year-olds tested in 1961, 1966, and 1980. 1985:02.

Lybeck, L. Research into science and mathematics education at Göteborg. 1985:03.

Lybeck, L., Strömdahl, H., & Tullberg, A. Students's conceptions of amount of substance and its SI-unit 1 mol. A subject didactic study. 1985:04.

Balke-Aurell, G. Testing testing methods. The Latin square design used in testing vocabulary by four methods. 1985:05.

Sandström, B. Studies of the process of innovation in the comprehensive school. 1986:01.

Lybeck, L. & Asplund Carlsson, M. Supervision of doctoral students. A case study. 1986:02.

Oscarson, M. Native and non-native performance on a national test in English for Swedish students. A validation study. 1986:03.

SKRIFTER FRÅN AVDELNINGEN FÖR SPRÅKPEDAGOGIK
GÖTEBORGS UNIVERSITET

Beställes från Institutionen för pedagogik, Göteborgs universitet, Box 1010, 431 26 MÖLNDAL

1. Oscarson, M. Engelska och svenska elevers prestationer på ett centralt prov i engelska. En valideringS- och utvärderingsstudie. Rapport nr 1986:02.

2. Hellekant, J. Franska elever gör ett svenskt franskprov. Ett försök till validering av det centrala provet i franska 1985. Rapport nr 1986:05

3. af Ekenstam, N-H. Tyska elever och svenska tyskprov. Hur klarar tyska gymnasister våra centrala prov i tyska? Rapport nr 1986:06.

4. Oscarson, M. Native and Non-Native Performance on a National Test in English for Swedish Students: A Validation Study. Report No. 1986:03.

5. von Elek, T. Invandrares språkutveckling under SFI-kurser. Rapport från ett kartläggningsprojekt inom AMU. Rapport nr 1986:13.

6. Lindblad, T. Betyg och Centrala Prov i Engelska, Tyska och Franska. Rapport nr 1986:14.

VASASTADENS BOKBINDERI AB
GÖTEBORG 1987