

THE SWEDISH SCHOLASTIC ASSESSMENT TEST (SweSAT)

Development, Results and Experiences

Christina Stage
Gunilla Ögren

EM No 49, 2004



ISSN 1103-2685

PREFACE

In this publication the Swedish Scholastic Assessment Test (SweSAT) is presented. Since spring 1977 SweSAT has been used as an instrument for selection to higher education. This publication is aimed at both those who would like a general description of the SweSAT and how it is used, and those who know more about test issues and would like to form a more detailed opinion about the SweSAT and the research that is done in connection to the test.

In Chapter 1 a short background is given and in Chapter 2 the development process and the content of the various subtests is described. In Chapter 3 tests for disabled groups are described, and in Chapter 4 the test results of different groups are presented. In Chapter 5 some measurement facts of the test are given, and in chapter 6 parts of the research activities connected with the test are described. Finally some opinions about the test are summarized.

The Agency for Higher Education is the responsible authority for the SweSAT and by direction of the Agency for Higher Education four of the five subtests of the SweSAT (WORD, DS, READ, and DTM) are developed at the Department of Educational Measurement at Umeå University. The fifth subtest (ERC) is constructed at the Department of Education at Göteborg University. The Department of Educational Measurement is also responsible for registration, scoring and standardization of the SweSAT.

Umeå, November 2004

Christina Stage
Scientific Leader of SweSAT

Gunilla Ögren
Project Manager

<http://www.umu.se/edmeas/hprov/english/>

1. INTRODUCTION	1
2. TEST DEVELOPMENT.....	3
2.1 SUBTESTS AND TEST CONTENT	3
2.1.1. <i>The Subtest WORD (vocabulary)</i>	4
2.1.2. <i>The subtest DS (data sufficiency)</i>	4
2.1.3. <i>The subtest READ (Swedish reading comprehension)</i>	5
2.1.4. <i>The subtest DTM (interpretation of diagrams, tables and maps)</i>	6
2.1.5. <i>The subtest ERC (English reading comprehension)</i>	6
2.2 PARALLEL TEST VERSIONS	7
2.3. SCORING AND STANDARDIZATION	8
3. TESTS FOR DISABLED GROUPS	9
4. TEST RESULTS OF DIFFERENT GROUPS.....	9
4.1 EDUCATIONAL BACKGROUND	9
4.2 AGE	10
4.3 MALES AND FEMALES	11
4.4 SOCIAL BACKGROUND	12
5. FACTS REGARDING MEASUREMENT TECHNIQUE	13
5.1 ITEM ANALYSIS	13
5.2 RELIABILITY	13
5.4 VALIDITY	15
6. DEVELOPMENT WORK.....	16
6.1 DIFFERENCES IN RESULTS BETWEEN MALES AND FEMALES	16
6.2 REPEATED TEST-TAKING	17
7. CONCLUDING REMARKS	18
REFERENCES	20

1. INTRODUCTION

Since spring 1977 the SweSAT has been used as an instrument for selection to higher education. Up until the spring of 1991 the test could be used by a limited group of applicants only, namely those who were qualified for higher education by age and work experience, i.e. those who were at least twenty-five years old and had at least four years of work experience. In 1991 new rules for admittance to higher education came into effect. These new rules mean that since the autumn of 1991 the selection for places for all applicants takes place either through grades from upper secondary school or through results on the SweSAT. The SweSAT was thereby given much more importance when it comes to admittance than previously.

The SweSAT is given twice a year, once in spring for admittance the following autumn, and once in autumn for admittance the following spring. The enrolment fee is SEK 350. For each test occasion a completely new test is constructed. The test, which is optional, is always carried out on a Saturday. The test scores are valid for five years, and it is allowed to take the test several times. The best valid result of the applicant is used for admittance.

The SweSAT is a norm-referenced test whose primary aim is to assess the test-takers' general aptitude for studies. The test should, as fairly as possible, rank the applicants with respect to expected success in higher education. Other requirements placed on the test are:

- The test should be in line with the goals and content of higher education and be relevant to the entire higher education sector.
- The test takers should regard the test as meaningful and suitable for determining entry to higher education.
- It should not be possible to improve individual results by mechanical practice or by learning certain strategies.
- The requirements for comprehensive recruiting shall be taken into consideration when constructing the test so that nobody is treated unfairly due to, e.g., their sex or social group.
- It should be possible to mark the test quickly, cheaply and objectively.

When the SweSAT was introduced in 1977 it consisted of six subtests. It has since then undergone some changes. In Table 1 the changes in the composition of the test are presented, as well as when the changes were made.

Table 1. Changes of the test, from the first version of the test up to the current. Subtests and number of items.

Subtest	1977-1979	1980-1991	1992-1995	1996-
WORD	30	30	30	40
DS	20	20	20	22
READ	30	24	24	20
DTM	20	20	20	20
STECH	20	20	-	-
GI	30	30	30	-
ERC	-	-	24	20
Totalt	150	144	148	122

Explanations for the subtest abbreviations

WORD Vocabulary

DS Data Sufficiency

READ Swedish reading comprehension

DTM Interpretation of diagrams, tables, and maps

GI General information

STECH Study technique

ERC English reading comprehension

In the spring of 1996 the SweSAT underwent fairly substantial changes. Since then SweSAT has consisted of 122 multiple-choice items divided into five subtests. The reason for these changes was that the pre-testing of items could take place in connection with the regular tests.

Up to 1995, pre-testing of items was carried out twice a year on third-year students at the theoretical programmes in upper secondary school. The model used had been criticized, especially on two accounts: the groups of test-takers were too small to give stable item characteristics and since the students knew that the results on the tests were of no personal importance, their motivation may have been lacking. Furthermore, the Swedish upper secondary school system was being reorganized and permission for pre-testing in the schools would become more difficult to obtain in the new system. This made it necessary to change the method for pre-testing the items for the SweSAT.

In connection with the changes of the pre-testing model the test day was divided into five test sessions of 50 minutes each (see Stage, 1995 or Ögren et al. 1995) of which one session consists of pre-test items. The subtests DS, READ and DTM each make up a session while the subtests ERC and WORD together form a session. The fifth test session contains newly constructed pre-test items. The order of the subtests varies between test occa-

sions. This pre-testing model means that all test-takers have to take two versions of one of the subtest sets. The set of pre-test items is not included in the test-taker's result.

2. TEST DEVELOPMENT

It takes about two years to produce a complete SweSAT, from the first draft of items to the final printed version of the test. Each item is pre-tested at least once among approximately 2,000 test-takers before it is included in a final version. The purpose of these try-outs or trials is to collect data regarding the functioning and difficulty of individual items. Following the trials, statistical, logical and content-oriented analyses are performed, which often result in modifications. For example, distractors that have appeared so unlikely that no one has chosen them are eliminated, while options that have seemed too likely although they are intended to be wrong are rephrased. Another purpose with the trials is to make sure that the various versions of the SweSAT used at different test occasions are equivalent regarding difficulty level and content. Test results are valid for five years, and those who apply for admission to higher education must be able to compete on equal terms regardless of when they have obtained their results.

After the trials, the items are grouped into subtest sets according to certain rules: certain types of items, certain subject areas etc. must be represented in the same way in each test. In addition, the various subject areas in the complete test must be represented in a way, which corresponds to subject contents in higher education. During this process the test items are reviewed and scrutinized several times by experts in various fields, such as subject matters, language, and testing. Each item included in the final test version has been checked and approved by up to ten people.

2.1 Subtests and test content

For all subtests the content must be varied and cover different subject areas in order not to give certain groups of test-takers unfair advantages or disadvantages. In order to achieve a suitable spread five subject areas, intended to reflect the different sectors of higher education, should be represented in each subtest. These sectors are: technology and science; administration, economics and society; health care; culture and information; and education and professions.

The items in the SweSAT should be of such a nature that they require effort at different cognitive levels. Generally speaking, the items should cover the three cognitive levels: recollection, understanding, deduction, and evaluation.

Apart from covering the five subject areas and requiring efforts at different levels the subtests also have to fulfil a number of requirements that are specific to each subtest.

2.1.1. The Subtest WORD (vocabulary)

The subtest WORD contains 40 items to be done in 15 minutes. The test is intended to assess understanding of words and concepts. Tests of this kind have proven to be good instruments when it comes to predicting study success.

As mentioned previously the different sectors of higher education are supposed to be represented in each subtest. For the subtest WORD this division cannot be completely unambiguous as most words can be regarded as "general" words.

WORD includes both Swedish and loan-words, especially words that are expected to be part of scientific texts. Each WORD-test contains 19 nouns, 11 verbs, 8 adjectives and 2 adverbs. The distribution between these parts of speech is based on Sture Allén's comprehensive study of frequencies of use in modern newspapers in *Tiotusen i topp* (1972).

The WORD-items are either synonyms or hyponyms. Synonyms are words that have the same or almost the same meaning, e.g. *angry* and *cross*. Hyponyms are words that denote a subset of the class of phenomena that the other word signifies, e.g. *tulip* and *flower*. Most items belong to the synonym group, but each test also contains some hyponyms.

The main words in WORD must not be so obvious that the entire group of test-takers know them, or so specific that only a select few have heard of them. Five options are constructed for each item. Examples of the subtest can be found in Appendix 1.

2.1.2. The subtest DS (data sufficiency)

The subtest DS contains 22 items that are intended to measure quantitative reasoning-ability. The test time is 50 minutes. The task presented to the test-takers is to judge whether there is enough information to solve a certain problem in an unambiguous way. The item format is fixed, which means that the options are identical for all items of the tests. Examples of the subtest can be found in Appendix 2.

The test presupposes knowledge of certain basic quantitative concepts and units defined by the mathematics curriculum up to and including course A¹ in upper secondary course, i.e.:

Arithmetic: number concepts, real numbers, growth factors, percentages, permillages, parts per million (ppm), index, prefix, powers with integer indices

Algebra: expressions and formulas, linear equations, simple power functions

Geometry: circumference, area, volume, angles, sine and cosine

Statistics: measures of central tendency and variation, probability

Calculus: graphic descriptions of linear equations, exponential functions

2.1.3. The subtest READ (Swedish reading comprehension)

This test is intended to measure Swedish reading comprehension in a wide sense. The test-takers have 50 minutes to read five texts and answer four questions on each text, i.e. reply to a total of 20 items. The texts cover a range of subjects and linguistic types, with an emphasis on popular science material, opinion pieces, reviews and similar material.

The five texts should, to as large extent as possible, cover the five sectors of higher education. The texts are taken from topical magazines, newspapers and books. The idea is that these sources should correspond to the type of written material that university students are expected to encounter and handle in their studies.

A reasonable linguistic variation is also required: factual text (studies, scientific reports, etc.), argumentative language (opinion pieces, letters to the editor, etc.), official jargon (legal texts, instructions, proclamations, etc.) and, to a certain extent, prose (fictional material). To present pure examples of texts of one kind or the other is, however, not an aim. The linguistic variation must also reflect an existing reality where different linguistic patterns overlap.

The questions about the texts are at different cognitive levels, with a prevalence of questions that in some way require understanding and characterization of the entire text and an ability to come to reasonable conclusions from

¹ Course A is a course in mathematics, which is compulsory for all students in upper secondary school, regardless of study programme.

the text material. Four options are constructed for each item. Examples of the subtest can be found in Appendix 3.

2.1.4. The subtest DTM (interpretation of diagrams, tables and maps)

This test is intended to measure the ability to interpret information in diagrams, tables and maps and other graphic presentations. Textbooks and teaching materials often contain such graphic presentations, which students need to be able to retrieve relevant information from.

The subtest DTM consists of 10 sets of figures and 20 questions. For each set of figures there are two questions with five options each. The test time is 50 minutes. The only aid allowed at the test session is a straight ruler.

Types of figures that may occur are, for instance:

- tables
- bar charts, histograms
- step functions, cumulative frequency graphs, curves
- pie charts, venn diagrams
- maps
- other (e.g. flow charts, time lines)

The figures are taken from both current and historical sources. Since the material has been removed from its original context it sometimes has to be supplemented with explanations of certain concepts or circumstances. However, the test developers strive to keep adaptations and changes to a minimum.

The items are aimed at covering the essential content of a set of figures, and questions to a certain set of figures are of as varying types as possible. The procedure for solving the items is varied so that different strategies must be used. The solution strategies usually required for DTM-items are straight reading, simple calculation, more advanced calculations and estimation. Some items require a combination of different strategies. Examples of the subtest can be found in Appendix 4.

2.1.5. The subtest ERC (English reading comprehension)

The subtest ERC consists of 20 items to be completed in 35 minutes. All items in this subtest are in the form of questions with four options connected to a given text material. The test starts with two longer texts (no more than

an A4-page per text), then some shorter texts follow, with one, sometimes two items per text. The test is concluded with a fill-in-the-blanks test with approximately five items where the test-taker has to choose the most suitable word for the context in different parts of the text, using a somewhat longer text as a starting point.

The ERC-items are aimed towards identifying essential information in the text rather than isolated details and towards central arguments and conclusions. Each ERC-test contains authentic texts within a broad spectrum of subjects, arts/social science as well as technology/natural science. The texts are usually taken from popular science magazines and other, more general, periodicals and magazines and sometimes also from books. Consequently it is not a question of extreme specialist texts requiring extensive previous knowledge, but texts that could be part of the syllabus in upper secondary school. Provided that the test-takers have the linguistic qualifications necessary to understand what they are reading, they can use the text to arrive at the right answer. Another basic principle is that the difficulty should be found in the text, not in the questions. This means that the choice of words in the questions should be kept at a relatively simple level. Examples of the subtest can be found in Appendix 5.

2.2 Parallel test versions

Since the SweSAT results are valid for five years it is necessary that the tests are comparable between test occasions.

This requirement in turn means that certain other requirements have to be fulfilled when designing the final test. On the one hand each subtest should be adjusted to a model for what the test should look like with respect to subject areas, cognitive level, content categories, item format, amount of information, informational value, types of figures and solution procedures. Individual items and thereby the entire subtest should have known, defined levels of difficulties and each item should discriminate in a satisfactory way and be balanced as to incorrect options. In Table 2 the results for the pre-test and for the regular tests since the spring of 2000 are presented.

Table 2. Results from pre-tests and regular tests for the subtests spring 2000 (00:A) until autumn 2003 (03:B).

Subtest	Pre-test	Regular test
WORD	21.8 – 22.2	21.6 – 22.2
DS	10.8 – 11.5	10.5 – 11.8
READ	11.4 – 12.0	11.9 – 12.9
DTM	10.7 – 10.9	10.2 – 11.5
ERC	10.6 – 11.0	11.3 – 12.1
Total	66.0 – 66.8	66.5 – 69.0

2.3. Scoring and standardization

Ever since the test was first used in 1977 the test has been marked in such a way that each correct answer has resulted in one raw-score point, the raw-score points for all the subtests have been summed up and the sum of the raw-score points have been transferred into a standardized score.

The main aim of the transference of raw-score points to a standardized score is to make the scores between different test occasions comparable, i.e. it should be just as easy or difficult to get a certain standardized score on one test occasion as on another. Since the period of validity has been extended from two to five years this similarity is therefore important over an even longer period of time than previously. This means that the standardization of the SweSAT results has become an even more delicate task, since the standardized score is to be comparable for ten test occasions.

The test designers' goal is to produce parallel versions of each subtest. These pre-equating efforts imply that the distribution of item content has to be the same at different test occasions both within the subtests and for the test as a whole. The distribution of the items' level of difficulty must be the same both within the subtests and for the test as a whole.

The standardization or post-equating takes place by weighing together different comparisons. First the results of all the test-takers are studied, but since the composition of the group of test-takers is not allowed to affect the possibility for a certain standardized score the results from two reference groups are also taken into consideration. The strategy used for standardization is to find the standardized values that best distribute the test-takers (in all three groups) over different standardized scores so that the correspondence to previous distributions will be as good as possible. In order to do this the equi-percentile method is used (see e.g. Holland & Rubin, 1982).

The standardized score is on a scale from 0.0 to 2.0, where 2.0 is the top score. For educational purposes standardization using Item Response Theory (IRT) has also been carried out.

3. TESTS FOR DISABLED GROUPS

Since autumn 1999 it is possible for people with dyslexia and people who are visually handicapped to take specially adapted tests. The adaptation for people with dyslexia entails taking the test with an extended time limit. As of 2002 mildly visually impaired have also been given the opportunity to take the test with an extended time limit. The test time has been increased by fifty percent per subtest and these groups do not have to do any pre-testing items.

For people who are severely visually impaired a special test has been assembled and transferred to Braille and audio cassette. This test contains 112 items and consists of 24 ERC-items, 30 DS-items, 18 READ-items (3 texts and 6 items for each text), and 40 WORD-items, i.e. for certain subtests the number of items has been increased and for some the number has been decreased. The test is also adjusted so that the test time is longer per item compared with the regular test. The subtest DTM has been excluded since there are no practical possibilities of transferring that subtest to Braille and audio cassette.

4. TEST RESULTS OF DIFFERENT GROUPS

After each regular test occasion the items are analyzed statistically and assessed. This analysis is mainly based on classical test theory. Not only individual items are analyzed but also the test-takers' results by educational background, age and sex.

4.1 Educational background

A comparison of results on the SweSAT of test-takers with different educational background shows that the longer the education the better the result on the test, i.e. there is a positive correlation between test result and educational background. Those who have studied the longest have the best results, up to 25 raw-score points (out of 122 possible raw-score points) higher than the group, which has the shortest background education (up to two years upper secondary school). In figure 1 the mean standardized scores for test-takers with different educational backgrounds (finished education) are presented.

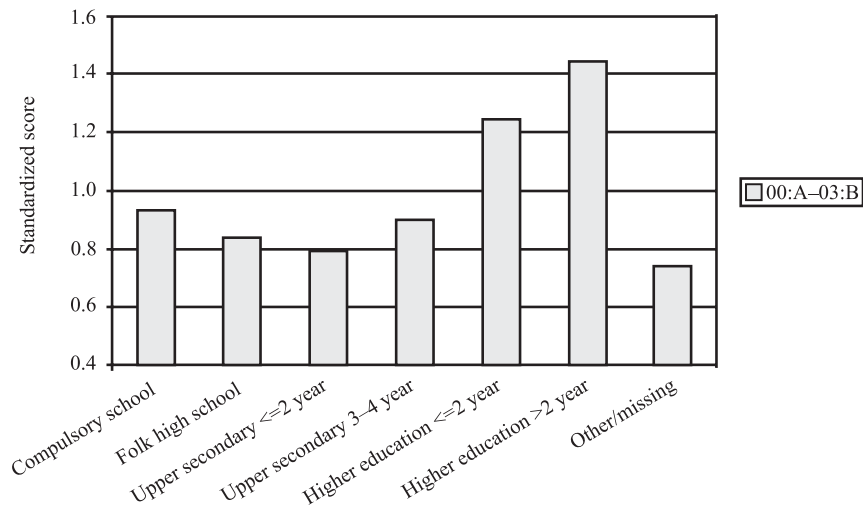


Figure 1. Mean standardized score for test-takers with different educational backgrounds 00:A-03:B.

In Figure 1 it may be seen that there is a positive correlation between educational background and test result. This is not surprising, considering that the test is intended to measure general aptitude for studies. Those who have studied for a longer period of time should, in all likelihood, be more used to studying and have a better study technique as well as more knowledge than those who have been studying for a shorter period of time and should therefore also achieve better results.

4.2 Age

Age has also proven to be important for the SweSAT result. Figure 2 the mean scores on the different subtests for test-takers of different age groups are shown for the tests in 2003.

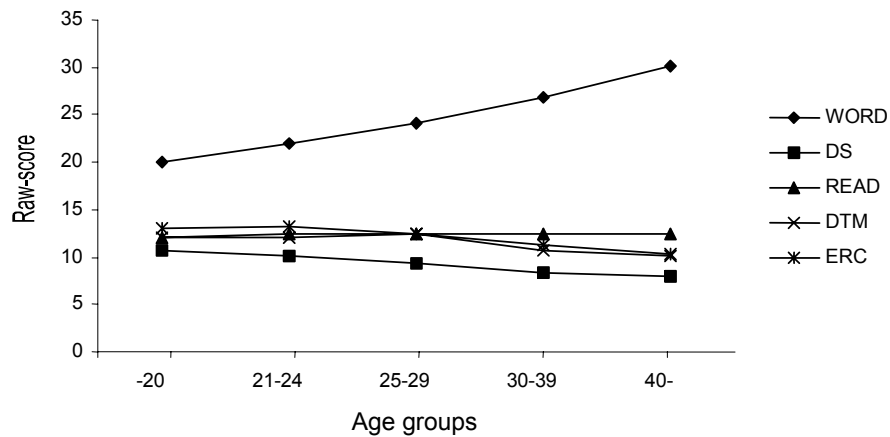


Figure 2. Mean test results (raw-score) on the different subtests for test-takers of different age groups in 2003.

From Figure 2 one can see that there is a positive correlation between the subtest WORD and age, i.e. the higher the age, the higher the mean result. For the subtests DS and DTM, on the other hand, there is a weak negative correlation with age, i.e. the higher the age the lower the result. The subtest ERC also shows a weak negative correlation with age. READ does not show any obvious tendencies as to correlation between results and age. The same relations between age and results on the different subtests have been found as long as the subtests have been used.

4.3 Males and females

Ever since the SweSAT was introduced there has been a very consistent difference between males and females in average results. Gender differences are not in any way unique to the SweSAT; on the contrary, gender differences are found on many standardized tests. In Sweden, however, the equality between the genders is an important issue and there is minimal acceptance of gender differences in test results regardless of the reasons for the differences. Generally the tests are blamed for being unfair and biased. The gender differences in results on SweSAT have been subject to extensive research with the purpose to describe, understand and explain the differences (see e.g. Stage 1985, 1990, 1993, Stage et al. 1995, Stage et al. 1998a, b). The two subtests that have the most quantitative elements, DTM and DS, account for the largest differences, over two raw-score points each. ERC accounts for approximately one raw-score point whereas the difference is

small for the verbal subtests, WORD and READ. Sometimes the result on the subtest WORD has been to the female's advantage (see Table 3).

Table 3. Differences in mean raw-score between males and females, for the different subtests 00:A-03:B. A negative score states that the difference is to the women's advantage whereas the others are to the men's advantage.

Subtest	00:A	00:B	01:A	01:B	02:A	02:B	03:A	03:B
WORD	-.63	-.39	-.48	.19	-.59	.28	-.62	.12
DS	2.48	1.74	1.95	2.04	2.13	1.60	1.86	2.11
READ	.25	.10	.05	.16	.09	.28	.05	.45
DTM	2.18	1.73	1.98	1.67	2.07	1.67	1.93	2.07
ERC	1.33	1.42	1.29	1.30	1.39	1.18	1.15	1.05
Total	5.60	4.61	4.78	5.36	5.10	5.02	5.61	5.80

4.4 Social background

The social background of the test-takers is also of importance to the results on the SweSAT. As a general rule test-takers from an upper-class background get better results than test-takers from a less privileged social background. This is the case on the test as a whole, on the subtest level as well as on item level. On the test as a whole the difference between upper class and working class are of about the same magnitude as the difference between males and females. On a subtest level the difference between males and females, as mentioned in section 4.3, is fairly large on the subtests DS and DTM, but almost non-existent on the subtests READ and WORD. When comparing social groups, on the other hand, it turns out that the differences are of the same size for all the subtests.

Another major difference between social groups is the percentage that takes the test. The upper class is over-represented among the test-takers whereas the working class is under-represented. This difference is accentuated further when it comes to the tendency to taking the test repeatedly, i.e. take several tests. The test-takers who use the opportunity to take more than one test usually come from a more privileged social background.

In conclusion, there is a clear connection between educational background and results on the SweSAT. There are also correlations, both positive and negative, between achievement on the subtests and age. The difference in mean results between males and females have, on the last few test occasions been 5 to 6 raw-score points. The differences described here, however, say nothing about results for individuals within each group.

In this line of work – construction and development of the SweSAT - it is necessary to continuously follow up and evaluate the test and the test results as well as other questions in connection to the test. Detailed accounts of results from the regular test occasions are continuously published in the *Pm-serien*, a series of reports published by the Department of Educational Measurement.

5. FACTS REGARDING MEASUREMENT TECHNIQUE

5.1 Item analysis

Analysis and evaluation of the tests' psychometric qualities, is first done at item level, thereafter on subtest level and finally on the test as a whole. Statistical analyses are carried out after each pre-testing of new items in order to evaluate each individual item with respect to whether it works well enough to be part of a regular SweSAT, as well as after every regular test to check that the items have worked as expected and in order to evaluate the pre-testing results.

The analysis of the pre-testing results include, among other things, evaluating the distractors (the incorrect options). At the same time it is examined whether those who have chosen the right option have a higher mean result on the subtest than the test-takers who have chosen one of the incorrect options. Furthermore it is examined whether there are any noteworthy group differences that could indicate that the item favours any specific group of test-takers over another. The biserial correlation of the items (r_{bis}), i.e. the item's correlation with the subtest result, is used for identifying anomalous items as well as for checking the homogeneity of the test. As a rule of thumb a limit of .30 is used, which means that, in order to be used as part of a regular test, an item must have an r_{bis} -value of at least .30. The proportion of correct answers (p-value) gives a measure of the level of difficulty and for each subtest it is stipulated how many items of each level of difficulty there should be and also the level of difficulty for the test as a whole.

From its inception the design and evaluation of the SweSAT have been based on *classical test theory*, but in later years parallel analyses have also been done based on *Item Response Theory*, IRT (see also chapter 6).

5.2 Reliability

The concept of reliability is central within test theory and refers to what extent the test results are consistent over different samples of items, test occasions, and groups of individuals. Reliability is consequently a measure of

how reliable the results are or how much trust can be placed in a test score. A reliability coefficient can be calculated in several ways, but always assumes a value between 0 and 1, where 1 is maximum reliability.

The reliability of the SweSAT is estimated by Cronbach's alpha (r_α), which also gives an indication of the homogeneity of the test. On the SweSAT as a whole, the reliability has been at least .91 and at most .93, i.e. it has been very stable and also quite high.

The reliability coefficients for the different subtests and for the test as a whole from spring 2000 until autumn 2003 are presented in table 4.

Table 4. Reliability-coefficients for the five subtests and for the test as a whole 00:A-03:B.

Test occasion	WORD	DS	READ	DTM	ERC	Tot
00:A	.85	.79	.70	.71	.82	.92
00:B	.84	.72	.69	.72	.80	.91
01:A	.85	.73	.71	.71	.80	.92
01:B	.85	.74	.69	.72	.78	.92
02:A	.85	.76	.75	.70	.80	.92
02:B	.86	.71	.70	.69	.79	.92
03:A	.86	.76	.74	.74	.81	.93
03:B	.81	.76	.73	.71	.78	.92

The reliability coefficients are of approximately the magnitude that can be expected on the type of test that the SweSAT constitutes. That the reliability is .92 means that for a test with the standard deviation 17 (which is the average standard deviation on the test after 1996) the error of measurement is 4.8.

An alternative way of determining reliability is to calculate the correlation between the results of a group of individuals on two parallel tests: parallel test reliability. A third way of determining the correlation between the results of a group of individuals on two subsequent test occasions: test-retest-reliability.

As was mentioned earlier one is allowed to take the SweSAT several times, which means that at each test occasion there are a number of people who

have results from previous tests. Correlations between test-takers' results on two test occasions are presented below (the number of people is given within brackets).

2003:A and B	.90	(6,549)
2002:A and B	.90	(5,789)
2003:A and 2002:A	.90	(5,191)
2003:B and 2002:B	.88	(3,607)

The correlations presented above constitute an under-estimation of the reliability since the time intervals between test occasions are six months or a year. During this time period it can be expected that a change has occurred in the test takers, which diminishes the correlations. The values presented above should therefore be considered a lower limit for reliability.

5.4 Validity

Since the SweSAT is an instrument for selection to higher education the test results are supposed to be correlated to study success, i.e. the test is supposed have good predictive validity or prognostic ability.

Prognosis of study success is an area that has been given a lot of attention in educational research. A number of Swedish and international studies present the ability of various selection instruments to predict study success. A somewhat simplistic summary of these studies is that the selection instruments that work the best are school grades and tests like SweSAT.

There are, however, a number of problems associated with studies of study success, one of which is the definition of study success. What is really meant by successful studies is by no means self-evident. The most common way of measuring study success has been to measure the students' grades after one or several years of their studies. This has worked rather well for certain international courses where there is regulated and differentiated grading. In Sweden, on the other hand, there are obvious problems. The grading in most universities and university colleges consist of, at most, three steps (fail, pass and pass with distinction), which means that there is not much variation within this criterion. Grading within higher education is not centrally regulated, which means that it is not comparable over different courses or institutions of higher education. A connected problem is whether good grades (whatever that means) necessarily is the best indicator of study success or whether there are other factors that are more important.

Another classic problem with prognostic studies is that one does not know anything about the applicants who are not admitted – would they have done

well or badly *if* they had been admitted? This is what is usually called *restriction of range* and which means that the correlation calculations result in a severe underestimation of the correlation that actually exists between the variables.

The problems of measuring study success and the difficulties of measuring predictive validity means that great efforts have been made to assess and evaluate the content validity of the test. Among the requirements for the test is that it should be *in line with the goals of higher education* and that it *should be perceived as meaningful and suitable for selection to higher education by the test-takers*.

The SweSAT is used for selection to various types of higher education courses and should therefore measure general qualifications for such studies. This manifests itself in the composition of subtests. The test assesses qualifications of both a verbal and non-verbal kind and the ability to take in information. Within each subtest the items are distributed over the five content areas that are broadly defined by the sectors within higher education.

6. DEVELOPMENT WORK

In connection with construction and evaluation of the SweSAT many questions arise. Ever since the start, the SweSAT has been the object of a number of studies. Examples of studies are evaluations of how different groups perform on the test as a whole, on the subtests, and on individual items. Other important issues are whether the existing subtests can be improved or replaced, whether the item types can or should be changed, and to what extent the different subtests are sensitive to the effects of practice. Other studies have been more basic in their nature, such as studies of the new test theory *Item Response Theory (IRT)* and the possibilities of applying this theory to the SweSAT (see Stage, 2003).

The studies that have been done regarding the SweSAT have continuously been presented in reports published by the Department of Educational Measurement. Up to and including 1984 this was done in the series *Spånor från Spint* and from 1985 in the *Pm-serien*. Since 1993 reports written in English have been published in the *Em-serien*.

6.1 Differences in results between males and females

Already at the pre-testing of the first trial versions of the SweSAT the problem of different mean results for males and females was brought to the fore. This issue has since been studied continuously. Different methods for *item*

bias have been tested, but, above all, attention has been focused on what areas and what content give rise to these differences.

It has been established that the two subtests with quantitative elements, i.e. DS and DTM, cause the greatest differences – over two raw-score points each. ERC normally causes a difference of about one raw-score point whereas READ and WORD result in approximately the same results for males and females. Studies on item level show that for all the items in the subtests DS and DTM the proportion of correct answers for females is somewhat lower than for males. The other subtests provide a different picture, namely that on certain items the males do better on average and on others the females do better on average; these differences (both ways) are the largest for items in the WORD subtest. However, on this subtest the differences even out and the mean result on the test is usually about the same for females and males. For ERC and READ the differences are smaller, but in ERC there are more items where males have a higher proportion of correct answers.

As was pointed out in Chapter 4, some of the differences in results can be explained by differences between the groups in terms of age and educational background. It has also been established that the mean differences between test-takers from different upper secondary school programmes are considerably larger than the mean differences between males and females.

Finally, it is important to remember that the variation within the groups is larger than the variation between the groups. This means that even if there are differences in mean result for males and females, the differences between the highest-achieving and lowest-achieving males are greater than the differences between males and females.

6.2 Repeated test-taking

For people who have taken the SweSAT several times it is always the best valid result that is taken into account at admission. This has led to a large number of people taking the test several times (see Table 5).

Table 5. Percentage of test-takers who on the test occasions 00:A-03:B took the test for the first, second, third or fourth time.

	Number of taken tests			
	1	2	3	4
00:A	59	28	10	3
00:B	63	25	9	3
01:A	60	27	10	3
01:B	68	22	8	2
02:A	67	24	7	2
02:B	70	21	7	2
03:A	69	23	6	2
03:B	70	21	7	2

Approximately 40 per cent of the test-takers at each test occasion has taken the test at least once before. On the last five test occasions this percentage has decreased to approximately 30 per cent. This naturally gives rise to the question whether test-takers who take the test several times also improve their results. A number of studies of the effects of repeated test-taking have been made (Henriksson, 1991, 1993, 1995; Henriksson & Bränberg, 1992, 1994; Henriksson & Törnkvist, 2002). To sum up, it can be said that test-takers who take the test more than once improve their results. The improvement seems to be the largest between the first and second test occasion, 0.1 standardized score point. After the second test occasion the increase at group level is smaller and smaller for each additional test taken. It should, however, be pointed out that this is the case at group level, whereas in individual cases there may be fairly large improvements and also deteriorations.

7. CONCLUDING REMARKS

In chapter 6 some of the big areas of study have been presented. Above that a number of small studies in connection with the different subtests are continuously performed, i.e. effects of the time limit, the difference between four and five answer choices. These studies are reported in the earlier mentioned Pm or Em series.

www.edmeas.umu.se/publications

In 2002 the National Agency for Higher Education decided to conduct an expert appraisal of SweSAT in connection with its 25th anniversary (Högskoleverket 2002:22R). A panel of three international experts in the field of testing were asked to assess whether the test could be regarded as an effec-

tive selection instrument, and to give recommendations about future development. The three experts found the SweSAT to be "... *a high quality test with a solid research program.*" (p. 19) Among the common recommendations given by the three experts were to:

- investigate the possibility of adding more context-based measures of reasoning, explore the possibilities of giving diagnostic feedback for counselling and individual score interpretation
- consider adding a computer-scored writing assessment in conjunction with
- exploration of the use of computer-based tests.

In 2004 a National Commission (SOU 2004:29) submitted proposals for new rules for admission to higher education. Regarding SweSAT the suggestions were that the test should also in the future be an important selection instrument for higher education, since research reports on SweSAT show that on the whole it functions well and in accordance with the original intentions.

However, evaluations and research studies performed during the years also show, that some aspects of the test should be scrutinized. Furthermore the test should be adapted to the development within the education area and in the society as a whole. In the development of the test the international, especially the American, development within testing should be followed. For specific domains (e.g. medicine, technical education) other selection instruments, besides SweSAT and grades, should also be developed.

SweSAT should be adapted to:

- people, who have Swedish as a second language
- the basic ideas characterizing authentic measurements
- information and communication technology
- students with special needs

And further:

- an essay test should be added.
- a computerized test should be developed as an alternative to the paper and pencil test
- all students in the last grade in upper secondary school should get the opportunity to take one test free of charge.

As may be seen above, there are a lot of more or less realistic ideas about changes of the present test. It is obvious that research and development work in connection with the test is very important. It is also very probable that continuous changes will take place as a result of that work.

REFERENCES

- Allén, S. (1972). *Tiotusen i topp. Ordfrekvenser i tidningstext*. Stockholm: Almqvist & Wiksell.
- Henriksson, W. (1991). *Effekter av upprepat provdeltagande* (Pm nr 40). Umeå: Umeå universitet, Pedagogiska institutionen, Avdelningen för pedagogiska mätningar.
- Henriksson, W. (1993). *The problem of repeated test taking and the SweSAT* (Em No. 5). Umeå: University of Umeå, Department of Education, Division of Educational Measurement.
- Henriksson, W., & Bränberg, K. (1992). *Effekten av upprepat provtagande. En studie av poängförändringar från första till andra provgenomförandet* (Pm nr 70). Umeå: Umeå universitet, Enheten för pedagogiska mätningar.
- Henriksson, W., & Bränberg, K. (1994). The effects of practice on the Swedish Scholastic Aptitude Test. *Scandinavian Journal of Educational Research*, 38(2), 129-148.
- Henriksson, W. (1995). *Repeated test taking and the SweSAT* (Em No. 13). Umeå: Umeå University, Department of Educational Measurement.
- Henriksson, W. & Törnkvist, B. (2002). *The Effects of Repeated Test Taking in Relation to the Test Taker and the Rules for Selection to Higher Education in Sweden* (Em No. 41). Umeå: Umeå University, Department of Educational Measurement.
- Holland, P.W., & Rubin, D. (1982). *Test Equating*. New York: Academic Press, Inc.
- The National Agency for Higher Education (2002). *A 25-year Testing program. Current status and future development*. Stockholm: The National Agency for Higher Education 2002:22 R.
- SOU 2004:29. Three Routes to the Open University. Official Report from the Commission for Admittance to Higher Education. Stockholm: Fritzes.
- Stage, C. (1985). *Gruppskillnader i provresultat*. Uppgiftsinnehållets betydelse för resultatskillnader mellan män och kvinnor på prov i ordkunskap och allmänorientering (akademisk avhandling för doktorsexamen). Umeå: Umeå universitet, Pedagogiska institutionen.
- Stage, C. (1990). *Könsskillnader i resultat på högskoleprovet våren 1990* (Pm nr 42). Umeå: Umeå universitet, Pedagogiska institutionen, Gruppen för pedagogiska mätningar.

- Stage, C. (1993). Gender Differences on the SweSAT. A Review of Studies since 1975. (Em No. 7) Umeå: Umeå University, Department of Educational Measurement.
- Stage, C. (1995). *Utprovning av provuppgifter* (Pm nr 106). Umeå: Umeå universitet, Enheten för pedagogiska mätningar.
- Stage, C. & Jarl, C. (1995). *Högskoleprovet hösten 1995. Provdeltagargruppens sammansättning och resultat* (Pm nr 110). Umeå: Umeå universitet, Enheten för pedagogiska mätningar.
- Stage, C., Ögren, G. & Konradsson, L. (1998a). *Högskoleprovet våren 1998. Provdeltagargruppens sammansättning och resultat* (Pm nr 139). Umeå: Umeå universitet, Enheten för pedagogiska mätningar.
- Stage, C., Ögren, G. & Konradsson, L. (1998b). *Högskoleprovet hösten 1998. Provdeltagargruppens sammansättning och resultat* (Pm nr 147). Umeå: Umeå universitet, Enheten för pedagogiska mätningar.
- Stage, C. (2003). *Classical Test Theory or Item Response Theory: The Swedish Experience*. (Em No. 42). Umeå: Umeå University, Department of Educational Measurement.
- Ögren, G. (2004). *Två, tre, fyra eller fem*. Effekter av förändringar av antalet uppgifter och antalet svarsförslag i ett av högskoleprovets delprov. (Avhandling för licentiatexamen). Umeå: Umeå universitet, Pedagogiska institutionen.
- Ögren, G., Stage, C., Åström, K., & Lexelius, A. (1995). *Försök inför förändring av högskoleprovets utprovningsrutiner* (Pm nr 107). Umeå: Umeå universitet, Enheten för pedagogiska mätningar.

The subtest WORD (Vocabulary)

1. **pipette**

- A magnifying glass
- B fine tube
- C clamp
- D small bottle
- E needlepoint

2. **proclaim**

- A clarify
- B guide
- C announce
- D scrutinize
- E decide

3. **invalidate**

- A endorse
 - B cancel
 - C schedule
 - D write off
 - E invoice
-

4. **abstraction**

- A major difference
- B liberty
- C conceptualization
- D incongruity
- E pointless task

5. **incarnate**

- A conquer
- B embody
- C illuminate
- D achieve
- E commence

6. **exposé**

- A lively discussion
 - B public commitment
 - C necessary condition
 - D strict interpretation
 - E public disclosure
-

7. **lucrative**

- A long-winded
- B appealing
- C complicated
- D profitable
- E entertaining

8. **polemic**

- A preventive treatment
- B remarkable idea
- C ability to co-operate
- D bitter argument
- E educational theory

9. **thereby**

- A after
 - B before
 - C without
 - D until
 - E because
-

Answers key:

1 B, 2 C, 3 B, 4 C, 5 B, 6 E, 7 D, 8 D, 9 E

The subtest DS (Data sufficiency)

1. The sum of three integers is 666. **What are the integers?**
- (1) The integers are n , $n+2$ and $n+4$, where n is an even integer.
 - (2) The least integer is 220 and the greatest integer is 224.

Sufficient information for solving the problem is given

- A in (1) but not in (2)
- B in (2) but not in (1)
- C in (1) together with (2)
- D in (1) and (2) separately
- E in neither of the two statements

2. A 380-kilometres trip was made with a car that had an almost full tank. **How many litres of petrol did the car use during this trip?**

- (1) After the trip there were 9 litres of petrol left in the tank.
- (2) After the trip the car was refilled with 36 litres of petrol. The tank was then full.

Sufficient information for solving the problem is given

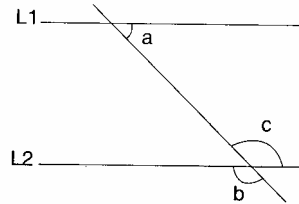
- A in (1) but not in (2)
- B in (2) but not in (1)
- C in (1) together with (2)
- D in (1) and (2) separately
- E in neither of the two statements

Appendix 2

3. L1 and L2 are two parallel lines. **What is the value of b?**

(1) $a=45^\circ$

(2) $c=135^\circ$



Sufficient information for solving the problem is given

- A in (1) but not in (2)
- B in (2) but not in (1)
- C in (1) together with (2)
- D in (1) and (2) separately
- E in neither of the two statements

Answers key:

1 D, 2 E, 3 D

The subtest READ (Swedish reading comprehension)

SARA LIDMAN AND GÖRAN TUNSTRÖM

Göran Tunström is one of the few authors who are able to combine naivety with intellect without creating cynical intruding sounds. His narrative is filled with approaches and embedments, developments and tickles, but his voice remains lonely, clear, and slightly brooding. With such novels as *Juloratoriet (Christmas Oratorio)* and *Tjuven (The Thief)* he reached a six-figure public; he already had the critics' interest. The public is not naive. Tunström can trust it to see his qualities, the vascular net of ideas and emotions under the round cheek of the narrative, because it can trust that he will never do anything at their expense. There is an episode in *Skimmer (Shimmer)* in which a father is advising his son in love. He asks him to note that the girl's body is hers and that it is his duty to let her experience it. Since all the text in a novel hangs together and as a whole expresses the novel's concept, we can imagine that this also applies to the author and his readers.

If we say that Sara Lidman brings to mind a figurehead, we convey something unintentional. She was, after all, a leading figure in the Swedish left during the 1960s and 1970s. But the image reappears as soon as one has rejected it. Her literature, her voice, project a figurehead on an old sailing ship, a vessel that can ply the waves and winds, in red, blue, and gold, who with eyes open and back straight leads her boat through the dusk and fog. Sara Lidman is an author who masterfully decides her own rhythm and her own pronouns. At the end of the seventies, she achieved the integration of a poetic provincialism, which had made her famous, and an internationalism in documentary form, which had made her controversial. It emerged in a suite of novels about how the railroad, how the modern world, came to a village in Norrland. When the rails bite through the marshes, they are an excellent backbone for the one who sits down to document and ponder the influence of modernity on human life in a landscape that is rich in resistance. Man was modern before technology, writes Lidman. He was modern in his ambivalence, in his knowledge of the unforeseeable, in his naive satisfaction with things modern, and in his traditionalism. *Lifsens rot (Life's Root)*, which is, and is not, yet another part of the Railroad Suite, contains a scene in which Rønnog, who in keeping with modern principles stopped nursing after one year despite the fact that her breasts were bursting with milk, is milking herself out in a barn. Suddenly she feels that someone is looking at her. A little boy is looking through a crack in the wall; his place at his mother's breast was recently taken by small siblings. She embodies and absurdly reflects his longing. On his forehead sits a mosquito, bursting with blood. Here we once again have a cell of a novel that is related to all the other cells and to the entirety.

Lidman is a matchless psychologist in her writing, something that is the direct opposite of describing individu-

als as a logical system. "Psychologic" would also be an unbelievable result in human relationships.

Tunström creates individuals with characteristics, and their lives unfold according to their own curves. Still, one cannot call his descriptions psychological.

Shimmer has a formulation. It is about Thorsteinn, "no intellectual giant. He never developed any cosmogony, never crept into the jungles of psychology, but he was, in his madness, wise, and built his life around an instant, giving him the appearance of profound energy." Tunström's style is anything but reductionist; still these words demand extra attention; they depict an attitude that signifies an aesthetic.

Both Tunström and Lidman "returned" in autumn 1996 with new novels after a decade of "silence." They were decidedly "events" and were treated as such, as the quotation marks try so clumsily to comment on ironically. For the authors, a decade without a new novel may have been agony (Tunström did, however, publish two other books in the interval), but the world's worries rather bring to mind those of the egg producer when his hens do not keep to a regular laying schedule.

While Sara Lidman's *Life's Root* is an independent continuation of the Railroad Suite, and masterfully goes over to a feminine track, Göran Tunström's *Shimmer* marks the beginning of a new orientation about whose future the reader can only guess. He has relinquished his province, Sunne in Värmland. Actually, he has done that before, but not with such an absolute, no-comment absence. The novel is Icelandic; not an American in Paris, but Icelanders in Iceland. Furthermore, the thematic discussion in his work between father and son has sunk to the father's range. And finally, the awareness of death is different from that found in earlier books where it was a stimulus, a rising to life, or living the curious mirror life of longing. There is an introspection in the perception of death, a curiosity of resignation, at times accompanied by the welcoming cello voice that Death has in the works of Schubert. Halldór, the father in the novel, reaches a state where the door stands ajar between life and death. He becomes aware and in some ways enviable. This is an interesting conquest, detrimental to the novel as whole. The conflict-filled and rivalrous love story involving Halldór and his only son, Pétur, is beautiful. But it is marred by banal wisdom intended to summarize how "we humans" live. And at times, the tall tale – which is otherwise one of Tunström's strong points – becomes mechanical, like an enforced narrative when he does not want to commit himself to the more reduced format in what is actually conveyed.

Sara Lidman can also be inclined to sacrifice herself too much to the tall tale. In her case, it is less a matter of filling out the novel than it is of wanting to explain too much. In addition to being a novel in itself, *Life's Root* is a collec-

tion of lives in the village. In the middle of the book, the reader finds himself among quickly terminated new starters that never move together to form a collective story. But what keeps it all in place is the story of Rønnog, and of the language that tells about her, makes her free to fly out of her fictive self and surprise the author, who once again accepts her and gives her her freedom once more. Rønnog is a girl from an ultra-religious, poverty-stricken family in Norrbotten at the turn of the century. "The threat" follows her throughout her life. There is a graft in her being, an unformulated weakness that she is terrified will overwhelm her and drive her mad or make her irresponsible. This threat has several names; all have to do with lost control. She is given the opportunity to learn to become a dairymaid. She wears a starched white uniform, pasteurizes the milk and makes cheese, gets her own life and authority. She falls madly in love with Isak Mårten, the oldest son of the patriarch Didrik in the Railroad Suite, and ends up all the same as a housewife in a mortgaged cottage in the forest, in a family

that places little value on control. Except for that, everything is valued. It's that simple. Milk is holy. Animals are not man's equal, but have their own value, which is not necessarily lower.

Rønnog's rage. It comes from the fact that her ambivalence is the exact opposite of her dominating father-in-law Didrik's. He believes in progress (electricity and telephones), but has a docile, almost carefree attitude to life. She has lived through progress, but in her everyday life she wants to get away from it since it gives an illusion that one is not abandoned.

A favorite cliché in literary small talk is that provincialism is true universalism. As we know, it can also be the exact opposite. Tunström and Lidman each has his or her dialectical landscape and each is also among the most widely traveled Swedish authors. Only there do they resemble one another. Like all good artists, they are differentiated by their qualities.

Sigrid Combüchen

Questions on the text

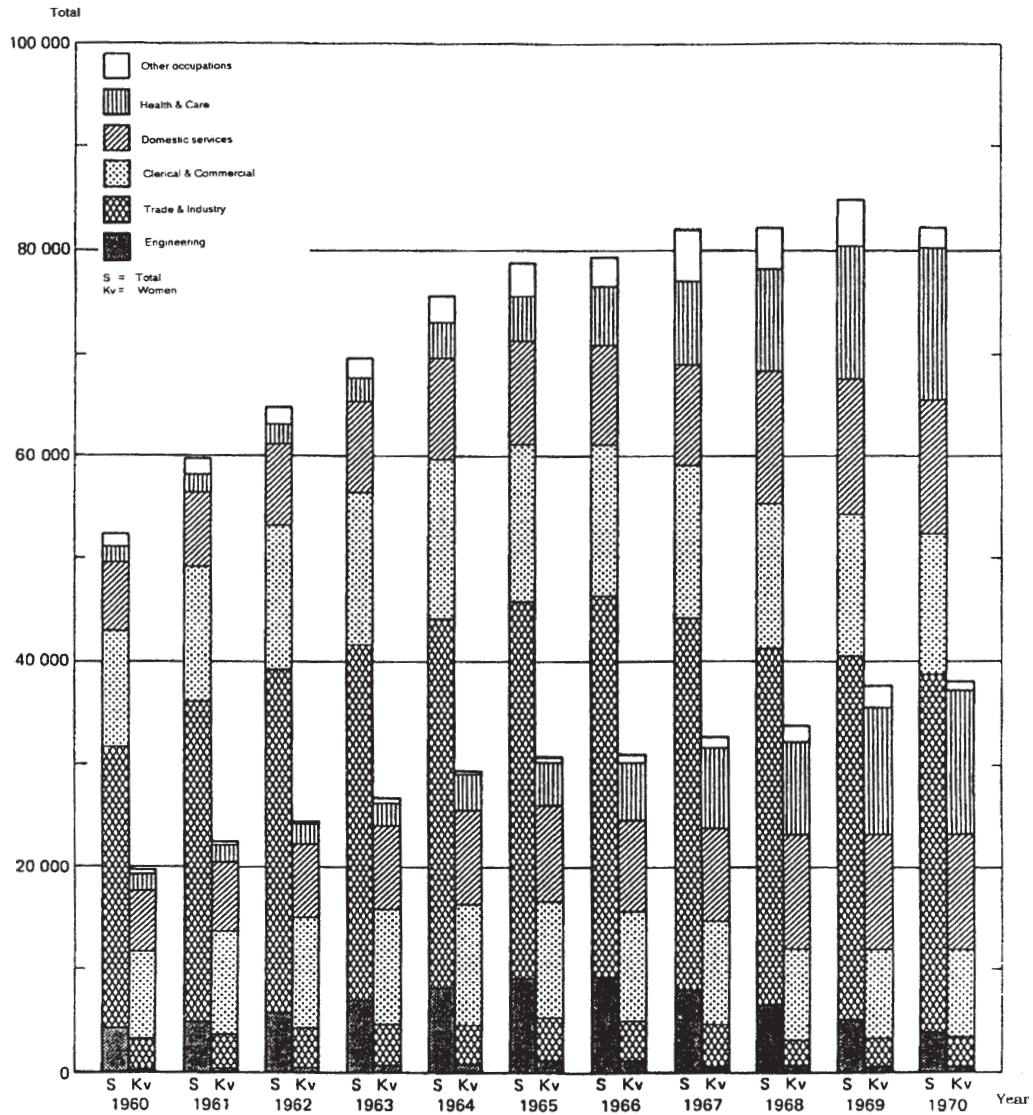
SARA LIDMAN AND GÖRAN TUNSTRÖM

- | | |
|---|--|
| <p>9. What do the writings of Lidman and Tunström have in common, according to the author of the text?</p> <p>A The ability to connect the modern with the traditional.</p> <p>B The combination of local roots and international outlook.</p> <p>C The ability to give voice to people otherwise neglected in society.</p> <p>D The combination of the unforeseen and the wellknown when it comes to the literary subject.</p> <p>10. How is the text written by Sigrid Combüchen best described?</p> <p>A Analysis</p> <p>B Piece of debate</p> <p>C Account</p> <p>D Personal portrait</p> <p>Answer keys:</p> | <p>11. What distinguishes the two referred novels from each other concerning their importance for Lidman's and Tunström's authorship respectively?</p> <p>A Tunström's novel breaks new ground while Lidman's is a continuation of her earlier novels.</p> <p>B Lidman's novel introduces wholly new characters while Tunström's clings to his old gallery of personas.</p> <p>C Tunström's novel tries out new aesthetical forms of expression while Lidman's refines her usual aesthetics.</p> <p>D Lidman's novel is open to the world while Tunström's clings to his provincialism.</p> <p>12. How is the attitude of Rønnog in <i>Life's Root</i> towards the modern development, according to the author of the text?</p> <p>A Naive</p> <p>B Confident</p> <p>C Uncertain</p> <p>D Resigned</p> |
|---|--|

9 B, 10 A, 11 A, 12 C

The subtest DTM (Interpretation of diagrams, tables and maps)

PUPILS IN VOCATIONAL COURSES 1960–1970



The total number of pupils and of them women pupils in full-time vocational school courses 1960–1970 distributed according to vocational sector.

1. Trade and industry was the vocational sector that had the largest number of full-time students during the period 1960–1970. **Which sector was second largest in 1960, and which in 1970?**
 - A Domestic services and then Clerical & Commercial
 - B Domestic services and then Health & Care
 - C Health & Care and then Domestic services
 - D Clerical & Commercial and then Domestic services
 - E Clerical & Commercial and then Health & Care

2. **What proportion of the total number of full-time pupils in vocational courses in 1967 consisted of women in the Clerical & Commercial vocational sector?**
 - A 12 per cent
 - B 18 per cent
 - C 31 per cent
 - D 40 per cent
 - E 67 per cent

Answer keys:

- 1 E
- 2 A

The subtest ERC (English Reading Comprehension)

Healthy Fats

If any nutrient could use an image makeover, it's fat. Often forgotten in our anti-fat frenzy is the fact that some fats are crucial for proper brain function. Among the good guys: the omega-3 fatty acids, which include linoleic acid—found in soybeans, canola oil, and nuts—as well as eicosapentaenoic acid (EPA) and docosahexaenoic acid (DHA), both plentiful in fish. Researchers have long known that infant brains require omega-3s, but now they say it appears these fats influence our behavior long after we've shed our diapers.

14 What *recent* information about omega-3s does the writer give?

- A They are good for our health
- B Adults' health may benefit from them
- C Children's health may benefit from them
- D They are found in fish

In Harmony At Last

In Mozart's time, most music was new music. Almost nobody listened to earlier composers besides a few antiquarians such as Baron van Swieten, who introduced Mozart to old J.S. Bach's scores.

Nowadays, **16**, "modern" music usually means 20th-century music, and a "contemporary" composer is anyone writing since 1945. Not even car-dealers and house-agents stretch words that far. Yet this wretched usage reflects a melancholy truth. For most of the century new classical music has had to struggle to win acceptance as the music of its time.

Happily, the gap between composers and audiences shows signs of narrowing at last. Professionals and listeners are adjusting to each other's **17**, Since Mozart, if not before, the best composers have always made technical and expressive originality their benchmark of success. Western music, as a rule, has been an innovating, not a conservative, tradition. Yet, audiences, who naturally want a musical language they can understand, have not always kept up.

The story of musical avant-gardes racing ahead of audiences is often told as if it were a peculiarly 20th-century problem. **18**, novelty has always caused trouble with the musical public, just as over-complication has with professionals. After the involved mathematics and ever-more tangled strands of Renaissance choral music, Monteverdi simplified things with a single expressive line. Then came competing lines and the intricacies of Bach. His style gave way in turn to simple tunes over a clear harmony which would shift at crucial points. For the next 100 years or so after Haydn and Mozart, composers explored ever subtler or more brutal shifts of harmony.

- 16**
- A if possible
 - B in addition
 - C for instance
 - D by contrast

- 17**
- A performances
 - B careers
 - C needs
 - D environment

- 18**
- A However
 - B Likewise
 - C Fortunately
 - D Consequently

Answer key: 14 16 17 18
 B D C A