Reports from the Department of Education and Educational Research

EVALUATION THROUGH FOLLOW-UP

GENDER DIFFERENCES ON THE SWEDISH SCHOLASTIC APTITUDE TEST

Sven-Eric Reuterberg

Report No. 1997:02
Department of Education and Educational Research
Göteborg University

EVALUATION
THROUGH
FOLLOW-UP

GENDER DIFFERENCES ON THE SWEDISH SCHOLASTIC APTITUDE TEST

Sven-Eric Reuterberg

Evaluation Through Follow-up is a research program aiming at a continous evaluation of the Swedish school system. The program was initiated by Statistics Sweden and the National Board of Education.

The project is financially supported by the Swedish Council for Planning and Coordination of Research.

ABSTRACT

Reuterberg, S-E. *GENDER DIFFERENCES ON THE SWEDISH SCHOLASTIC APTITUDE TEST.*

ISSN: 0282-2156

Number of pages: 35

Since 1977 the Swedish Scholastic Aptitude Test (SweSAT) has been used as an admission test for applicants to higher education in Sweden. Permanently the test results have shown gender differences in favour of male test takers and this has brought about an animated public debate. Very often the gender differences have been regarded as bias in the test. That is to say that the test has been unfair to female test takers.

However, before any conclusions as to fairness of the test can be drawn we have to take into account the fact that the test takers constitute self-selected groups and there is no reason to assume that male and female test takers should achieve equally well on the test. There may be gender differences in those abilities that are measured by the test.

The present study focuses on the gender differences in the SweSAT scores, when differences in test scores from the age of 13, marks from compulsory school and the test takers' upper-secondary education are taken into account.

According to the results the male test takers outperform the female test takers on all subtests and these gender differences cannot be explained by differences in test scores from the age of 13 or by differences in marks. When marks are taken into account, the gender differences often become even greater due to the fact that the females in most instances have the highest marks. Nor can the gender differences in SweSAT scores be explained by differences in upper-secondary education.

INTRODUCTION

Since 1977 the Swedish Scholastic Aptitude Test (SweSAT) has been used as an admission test for entrance into higher education. This was initially brought about by the fact that even persons without any leaving certificates from upper secondary education were eligible to enter certain programmes of higher education, and from the beginning the SweSAT was reserved for adult applicants who had at least four years of work experience. In 1991 the SweSAT became an alternative to the leaving certificates from upper secondary school as a selection instrument into higher education for all categories of applicants. These new admission rules resulted in an enormous increase in the number of test takers. During the 80's the number of test takers was constantly about 5.000 a year, but since 1991 the numbers have been between 100.000 and 150.000 (Ingerskog & Stage, 1993).

This study is based on a SweSAT version which contains six time limited subtests. These are:

Vocabulary, WORD, which measures understanding of words and concepts. The task is to identify which out of five words that has the same meaning as a given word.

Data Sufficiency, DS, which aims at measuring numerical reasoning.

Reading Comprehension, READ, which contains four texts and six multiple choice questions in relation to each text.

Interpretation of Diagrams, Tables and Maps, DTM, which consists of 10 collections of tables graphs and/or maps with two questions in relation to each such collection.

General Information, GI, which measures knowledge and information that a person may acquire in different contexts.

Study Techniques, STECH, which consists of a booklet with a number of texts about a subject matter area. The task is to find the answers in the booklet to 20 different questions.

In 1992 the STECH was replaced by a test of *English Reading Comprehension, ERC*. Due to new pre-testing routines the GI subtest was eliminated in spring 1996 which means that the SweSAT contains only five subtests at present.

All items are dichotomously scored and the total test score is the number of correctly answered items. For more comprehensive information on the SweSAT, its content and history the reader is referred to Wedman (1994).

The SweSAT shows gender differences in favour of male test takers and has done so constantly since it was introduced. Up to 1990 the gender differences amounted to an average of about 7 points (Stage, 1992) and in 1992, when STECH was replaced by ERC, the gender differences increased to 10 points which corresponds to about half a standard deviation unit (Ingerskog & Stage, 1993). According to Stage (1992) the gender differences have always been smallest on the WORD subtest, and on the whole the verbal subtests have shown moderate differences in favour of males. On the other hand, the greatest gender differences have always been found in the quantitative subtests, DS and DTM, which is in accordance with earlier results on mathematical tests (Hyde, Fennema & Lamon, 1990). In standardised mean differences the males surpass the females by 0.60 to 0.70 on these tests, while the differences are 0.20 to 0.25 on WORD and READ and about 0.40 on ERC. According to Stage (1992) the gender differences do not vary only between the subtests but also between test takers from different programmes of upper secondary education. The smallest differences are found among students of humanities and the greatest differences among students of economics. Stage also shows that the gender differences vary with age, being smaller among test takers in the ages of 25 to 39 than among younger and older test takers respectively. A similar trend is found also by Bränberg et al. (1990) in a study of the SweSAT tests administered in 1986 and 1987, that is to say when all testees were 25 years and older.

Gender differences in favour of male test takers are not only found in the SweSAT. The same is true also for the American counterpart, the Scholastic Assessment Test (SAT). According to Wilder and Powell (1989) males have traditionally surpassed the females by about half a standard deviation unit on the mathematical part of the SAT, but up to the early 70's the females obtained higher scores on the verbal part of the SAT. Since then, however, the males obtain higher scores even on the verbal sections.

In Sweden the gender differences attracted little attention in public debate as long as the test was taken only by adult applicants, but when the SweSAT was given a more important role as an admission test and was also taken by younger applicants an intense debate started. Often these gender differences were interpreted as bias in the test. SweSAT was biased in favour of male test takers, it was claimed. In order to reveal possible causes of this presumed bias, several studies were conducted focusing the impact of item content and format, testing time, item position, and problem solving strategies (Henriksson, Stage & Lexelius, 1986; Stage, 1987; Wester-Wedman, 1992a; 1992b; 1992c; Mäkitalo, 1996). However, so far these studies have not resulted in any changes of the SweSAT which have decreased the gender differences.

Of course, it is a rash conclusion to claim that gender differences in test scores, by themselves, indicate test bias (Cole & Moss, 1993). Many other causes are conceivable. There may, for instance, be gender differences in the abilities measured by the test. More than 20 years ago Maccoby and Jacklin (1974) regarded a greater verbal ability among girls, and a superiority among the boys in mathematical ability as two "fairly well established facts". Some later studies have found that these differences have declined over time (e.g. Feingold, 1988; Hyde & Linn, 1988; Hyde et al., 1990; Linn, 1991; Rosenthal & Rubin, 1982; Stumpf and Stanley, 1996; Cole, 1997). The evidence of declining gender differences are not conclusive, however. Cleary (1992), for instance, concludes that there are "no consistent trends over time" (p. 76) and Emanuelsson et al. (1993) studying nationally representative samples of 13 years old boys and girls found that the girls showed a more positive development than did the boys during the 70's on a verbal, a spatial and a reasoning test. During the 80's this trend was broken, however, and the girls lost their superiority on the verbal test and fell successively behind the boys on the other two tests.

When interpreting the gender differences in the SweSAT scores it must also be remembered that the test takers constitute a positively selected group. According to Cleary (1992) selectivity in the sample may have a substantial impact on gender differences. She found that the gender differences changed toward favouring males as the sample grew more positively selective. Thus, the fact that the SweSAT takers constitute a positively selected group may by itself be one contributing factor to the male superiority in SweSAT scores.

As to the SweSAT it is not only a question of selectivity. As shown by Mäkitalo (1994) and Mäkitalo and Reuterberg (1996), the male SweSAT takers constitute a more positively selected group out of all males than do the female test takers out of all females. Thus, it is also a question of differential selection to the SweSAT. In these studies the degree of differential selectivity was measured by test scores from the age of 13, standardised achievement test from grades 6, 8 and 9 in compulsory school and by school marks from grade 9 in compulsory school. If we make the assumption that there are no gender differences at all in the total group out of which the test takers are selected, then both the positive selection and the differential selection to the SweSAT can be expected to bring about gender differences in results.

As discussed by Reuterberg (1996) the selection effects and the differential selection effects, by themselves, do not indicate what differences we can expect in the SweSAT scores. Besides these effects we also have to take into consideration those differences that are found within the total sample. If, in the total sample, the males surpass the females, the various selection effects will reinforce the gender differences among the SweSAT takers. If, on the other hand, the females are superior to the males within the total sample, the two kinds of selection effects will reduce the gender differences among the test takers or even turn them into a difference in favour of males. Thus, the studies conducted by Mäkitalo and Reuterberg give important information on the processes which determine the composition of the group of test takers, but they do not indicate what gender differences we may expect to find in the SweSAT scores.

As has been stated repeatedly, results from statistically unrepresentative groups like the SweSAT takers do not lend themselves to any valid generalisations without relevant adjustments (Howe, 1985; Wainer, 1986a; 1986b; 1993). Nor can any conclusions be drawn concerning possible bias in the test from such self-selected groups, unless there are relevant variables available for making the necessary corrections.

Wainer and Steinberg (1992) used the type of course taken and the marks received in first-year college math courses as control variables for examining possible bias in the mathematical section of the Scholastic Assessment Test (SAT-M). Two different analyses were performed; one retrospective in which the SAT-M scores were compared between males and females with identical courses and equal marks, and one prospective in which the SAT-M score constituted the predictor of college math performance. Both analyses showed a gender difference in SAT-M scores in favour of male test takers. However, as pointed out by the authors, these results cannot be taken as a proof of bias in the test. The main problem is that the control variable was measured after the students had been selected to the courses and, therefore, differential selection mechanisms to the math courses might well be responsible for the gender differences in SAT-M scores.

Thus, Wainer's and Steinberg's study points out another important prerequisite concerning the control variables. They have to be measured before the groups have been the objects of any kind of selection processes. Consequently, the question of possible bias in admission tests cannot be answered unless longitudinal data are available. The present study is based on that kind of data and it aims at studying whether the gender differences in SweSAT scores in favour of males can be explained by corresponding differences in ability and achievement measured before any educational selection has occurred or by differences in the choice of upper secondary programme. If this is the case, the gender differences in SweSAT scores cannot be regarded as indicating bias in the test.

METHOD

Subjects

As mentioned in the preceding section, the question of bias in the SweSAT scores cannot be answered unless longitudinal data are available for a nationally representative sample of persons. Furthermore, the sample size must be large enough for giving reliable results also for those subgroups which are obtained when different control variables are introduced in the analyses and the data must also include relevant control variables which are not influenced by the criterion variable.

The Swedish longitudinal project called Evaluation Through Follow-up (ETF) satisfies all these requirements. This project has followed up nationally representative samples of pupils born in 1948, 1953, 1967, 1972, 1977 and 1982, respectively, from the age of 10 or 13 and all through the formal school system. A more comprehensive description of this project is given by Härnqvist, Emanuelsson, Reuterberg, and Svensson (1994). The cohort of interest in this study is the one born in 1972. The 9.000 persons in this cohort were in grade 3 of compulsory school in the spring of 1982 when the follow-up study began. Since the sample is drawn out of all individuals in that grade, their year of birth varies, but the great majority (95 per cent) were born in 1972.

The sample has been followed up to the age 20 with repeated data collections. The latest data, the SweSAT scores, are received from another large data base called 'BACE 72' including everyone born in 1972. Up to now, however, only the SweSAT scores from 1990 to 1992 are available in BACE 72.

A set of matched data are available for in all 8.728 individuals and about one person out of four has taken the SweSAT during the period mentioned above. However, in this study only two test occasions are analysed, those from the spring of 1991 and 1992, called 91A and 92A, respectively. The reason for studying only the spring test occasions is that more people take the SweSAT in the spring and it is assumed that the autumn testees are too few to give reliable results.

As shown by Table I, about 14 per cent of the total sample took the SweSAT in the spring of 1991 and nearly 11 per cent took the test one year later. Due to missing data, the group sizes are reduced so 1.030 persons remain for the analyses of the results from 91A and 784 for the results from 92A. These numbers correspond to 12 and 9 per cent, respectively, of the total sample. About 300 persons have taken the SweSAT on both these occasions.

Table I. Number of individuals in the sample, drop outs, and number of testees.

	Males	Females	Total
Total sample	4449	4279	8728
Testees 91A	559	656	1215
Missing control variables	60	76	136
No upper secondary education	16	33	49
Remaining for analyses of 91A	483	547	1030
Testees 92A	484	453	937
Missing control variables	54	53	107
No upper secondary education	24	22	46
Remaining for analyses of 92A	406	378	784

Variables

As mentioned in the introductory section the SweSAT version on which this study is based contains six subtests. For each of them a subtest score has been computed which corresponds to the number of correctly answered items. These subtest scores constitute separate variables and they are designated according to the subtest in question. By summing the subtest scores we obtained the total SweSAT score called *SweSAT*.

Two categories of control variables are used, namely test scores collected in grade 6 and marks from grade 9 of the compulsory school. Since there is no streaming in compulsory school during the first six school years, all individuals have had the same formal education when the tests were taken. Three different tests were administered in grade 6:

Opposites (Op) is a traditional test measuring verbal ability. It includes 40 multiple choice items and the task is to select one word out of four, which is the antonym of a given word.

Metal Folding (MF) measures spatial ability. In this test the task is to identify a three-dimensional figure among four flat pieces of metal with bending lines. The test contains 40 items.

Number Series (NS) measures reasoning ability. In each of the 40 items six numbers are given which are ordered according to a mathematical rule. The respondent's task is to detect the rule and to add the next two numbers in the series.

The scores on these three tests are summed into a total score *Testsum* which constitutes a measure of general intellectual ability.

In grade 9 all pupils receive marks in each school subject studied. These marks range from a highest value of 5 to a lowest of 1. For the whole population the marks should be normally distributed with a mean of 3 and a standard deviation of 1. In English and Mathematics the pupils have to choose between a general and an advanced course in grades 7 through 9. Therefore, the pupils in each of these courses constitute their own reference group, which implies that the marks are not comparable between the two course alternatives. In order to make them comparable correction factors have been estimated. These corrections imply that the marks from advanced course in Mathematics have been increased by a constant of 1.01 points (Reuterberg, 1994). The corresponding correction factor for the marks from the advanced course in English is 0.77.

The marks have been combined into five subject areas:

Language (Lang) which is the mean of the marks in Swedish and English.

Natural sciences (Mathsci) which is the mean of the marks in Mathematics, Physics, Chemistry and Biology.

Social sciences (Socsci) which is the mean mark for the following subjects: Geography, Civics, Religion, and History.

Practical-spatial (Pract) which is the mean mark for Drawing, Handicraft and Technology.

Finally, a variable called *School achievement (Sach)* has been computed and it is based on the mean of the marks for all school subjects studied.

Of these control variables MF, Socsci, and Pract have shown so low correlations with the SweSAT variables that they have been excluded from any further analyses. The remaining control variables have been tested in different constellations in order to find out those combinations of one test variable and one mark variable which have the highest predictive power in relation to each of the SweSAT variables. The results of these screenings are found in Appendix and they ended up with the following combinations:

SweSAT variable		Combination of control variables
SweSAT	<--------->	Testsum, Lang
WORD	<--------->	Op, Lang
READ	<--------->	Op, Lang
STECH	<--------->	Testsum, Lang
ERC	<--------->	Op, Lang
GI	<--------->	Op, Lang
DS	<--------->	Testsum, Mathsci
DTM	<--------->	Testsum, Mathsci

Thus, NS will be excluded from the forthcoming analyses in spite of the fact that it has shown a fairly high prediction value in combination with Lang and Mathsci, respectively, for DTM. However, the combination of Testsum and Mathsci showed an even higher predictive power in this case. Also Sach will be excluded since it has shown moderate predictive power throughout, and this is due to the fact that it includes some school subjects which are weak predictors for the SweSAT variables i.e. Sports, Handicraft and Drawing and Home economics.

As mentioned in the introduction, the gender differences in SweSAT scores differ from one upper secondary programme to another. Therefore, previous education has to be taken into consideration. Persons without any upper secondary education are excluded because very few of them have taken the SweSAT, so six educational groups remain:

Students from the 2 year programmes (2YEAR)
Students from the 3 year Humanities programme (H)
Students from the 3 year Economic programme (E)
Students from the 3 year Social science programme (S)
Students from the 3 year Natural science programme (N)
Students from the 3 or 4 year Technical programme (T)

Due to small group sizes all these categories cannot be separated, and therefore we have to work with the following three educational categories:

2YEAR
HSE
NT

Table II. Number of individuals in different educational groups.

	SweSAT 91			SweSAT 92		
	Males	Females	Total	Males	Females	Total
2YEAR	13	24	37	37	36	73
HSE	112	315	427	150	231	381
NT	358	208	566	219	111	330
Total	483	547	1030	406	378	784

Table II shows the typical gender distribution for the different educational groups with females in majority among those who have finished a program of humanities, economics and social science and males are in majority on the natural science and technical programmes. The 2YEAR group is the smallest one and it shows an even gender distribution among those who took the SweSAT in 1992, while there are more females than males among the 1991 test takers.

In Table III the means of the control variables are shown. These means differ quite a lot from those found for the total age group. In the first place, they are all quite high due to the fact that the SweSAT takers constitute a positively selected group. In the second place, the gender differences are affected by differential selection to the SweSAT implying that the male test takers constitute a more positively selected group out of all males than do the female test takers out of all females.

Within the total cohort there are no gender differences in Testsum and moderate Op differences in favour of the females (Mäkitalo, 1994, p. 16-17). Due to the differential selection effects to the SweSAT Table III shows a Testsum difference in favour of the male test takers and only small and inconsistent differences in Op.

As shown by Svensson (1971) females obtain higher marks in grade 6 than we could expect from their test scores and this gender difference in relative school achievement tends to increase from grade 6 to grade 9 (Härnqvist, 1993). Therefore, in the total age cohort, the females have obtained substantially higher scores than the males in Lang and moderately higher scores in Mathsci. Among the test takers there is still a difference in favour of the females in Lang, but when we turn to Mathsci we can see that the male test takers score somewhat higher. The last mentioned difference, however, is not consistent over the various educational groups.

Table III. Means for control variables in relation to gender and educational group.

Control variable	Educat. group	SweSAT 91		SweSAT 92	
		Males	Females	Males	Females
Testsum	2YEAR	72.62	72.25	70.16	68.00
	HSE	80.27	79.74	81.03	80.10
	NT	87.21	86.99	84.98	85.59
	All	**83.59**	**82.17**	**82.17**	**80.56**
Op	2YEAR	24.15	23.88	24.30	22.97
	HSE	27.14	27.13	26.78	27.41
	NT	27.67	28.50	27.24	27.78
	All	**27.48**	**27.51**	**26.80**	**27.10**
Lang	2YEAR	3.26	3.55	3.36	3.60
	HSE	3.99	4.33	4.00	4.26
	NT	4.20	4.56	4.08	4.46
	All	**4.12**	**4.38**	**3.98**	**4.26**
Mathsci	2YEAR	3.12	3.28	3.02	3.11
	HSE	3.71	3.74	3.69	3.55
	NT	4.20	4.26	4.10	4.22
	All	**4.06**	**3.92**	**3.85**	**3.70**

A comparison between the educational groups shows a general trend of NT having the highest means and the 2YEAR group the lowest. This pattern is to be expected in the light of the differential selection mechanisms to the upper secondary programmes with NT as the most positively selected and 2YEAR as the least selective group. We can also see a general trend of somewhat lower means for the group that took the SweSAT in 1992, which is natural in the light of the fact that this group also contains applicants who were not admitted to higher education in 1991.

Statistical method

The statistical method used is multiple regression analysis with the various SweSAT scores as the dependent variable and gender (SEX), marks, test scores and previous education as independent variables. The analyses will proceed in the following sequence:

Independent variables	Information given
SEX	The actual gender difference in SweSAT scores.
SEX + Mark	The gender difference in SweSAT scores, when differences in marks are controlled for.
SEX + Test	The gender difference in SweSAT scores, when the differences in test scores are controlled for.
SEX + Mark + Test	The gender difference in SweSAT scores, when differences in both marks and test scores are controlled for.

The results obtained by these analyses show the gender differences when no account is taken to differences in previous education.

In a second step the programme chosen in upper secondary school will be introduced together with the other independent variables in the same sequence. This step gives the same information as that in the first one, but separately for each educational category.

SEX and previous education are handled as dummy variables with females and persons with a 2 year upper secondary education as reference groups. This means that positive gender differences imply a higher mean for the male test takers. The differences found for the total group will be tested for statistical significance on the 5 per cent level, but there will be no statistical tests of the differences found within each educational group, due to the fact that the statistical method used gives t-values for the reference group only i.e. test takers from the 2 year secondary school programmes.

In order to make the gender differences comparable between the total SweSAT scores and the various subtests, they will be expressed as effect sizes (ES). This means that the raw score difference is divided by the pooled within-group standard deviation.

Besides the effect sizes R^2-values are also presented to show to what extent the variance in the dependent variable is explained by sex alone or by sex in combination with the control variables. Since these values are additive they show the contribution of each additional control variable to the explanation of the variance in the SweSAT variables studied.

RESULTS

The first gender differences to be analysed are those for the total SweSAT score, followed by the results for each separate subtest. As mentioned above the differences are expressed as effect sizes and positive values indicate that the males have the highest mean.

SweSAT

Table IV. Gender differences in SweSAT scores.

Control variable	1991 ES	1991 R^2	1992 ES	1992 R^2
None	.67*	.09	.53*	.05
Lang	.93*	.39	.82*	.37
Testsum	.51*	.46	.45*	.39
Lang + Testsum	.62*	.56	.67*	.51

* Significant at the 5%-level.

All gender differences in Table IV are significant at the 5%-level and in favour of male test takers. Among those who took the SweSAT in the spring of 1991 the actual gender difference amounts to nearly 0.7 standard deviation units. This corresponds to more than 10 raw score points. On the test occasion one year later the difference is somewhat smaller - about half a standard deviation unit. To some extent these differences are explained by differences in Testsum. When this variable is kept under control they decrease by about one tenth of a standard deviation unit, but they are still statistically significant. This means that if we select a group of male test takers and a group of female test takers with exactly the same test scores, then the males' SweSAT raw score mean would exceed that of the females by 7 or 8 points.

Taking gender differences in the marks in Swedish and English from compulsory school (Lang) into consideration, the picture becomes even more dramatic. Since the females have received higher marks we should expect them to get a higher SweSAT score than the males, but as we have already seen this is not the case. With the marks as control variables the gender differences in favour of males are increased to 0.9 and 0.8 standard deviation units respectively, or to 14 and 12 raw score points. When both control variables are taken into consideration simultaneously more than 50 per cent of the variance in SweSAT scores is explained and the gender differences fall between those when only one of the variables was controlled for.

The conclusion to be drawn on the basis of these analyses is that the gender differences in SweSAT scores cannot by far be explained by differences in ability as measured by tests in the age of 13 or by differences in school achievement as measured by marks from compulsory school.

In these analyses the assumption of parallel regression lines for males and females has been violated when Lang is taken as the control variable. In this case the SweSAT scores increase more rapidly for the females than for the males with increasing mark level. However, irrespective of mark level the males get the highest SweSAT scores. Thus, there is a general trend of decreasing gender differences in SweSAT scores with increasing mark level.

Table V. Gender differences in SweSAT scores by educational group.

Control	1991				1992			
variable	NT	HSE	2YEAR	R^2	NT	HSE	2YEAR	R^2
None	.40	.53	.70	*.19*	.24	.53	.53	*.24*
Lang	.73	.84	.97	*.42*	.58	.76	.74	*.43*
Testsum	.39	.51	.68	*.48*	.27	.49	.44	*.46*
Lang + Testsum	.61	.71	.88	*.56*	.50	.66	.61	*.54*

A comparison of the R^2-values in tables IV and V indicates that selections of course at the upper-secondary level can have some bearing on future SweSAT scores. The proportion of explained variance in the actual SweSAT scores increase from 9 per cent to 19 for the 1991 test and from 5 to 24 per cent for the 1992 test. However, this increase occurs mainly through differences between the educational groups in Lang and Testsum. When these two variables are included in the regression equations, the R^2-values amount to about the same values as when previous education was not accounted for. Thus, the contribution of the upper secondary programme chosen to the proportion of explained variance in SweSAT scores does not occur via school knowledge acquired on that level, but via differential selection to the various programmes.

As Table V shows, gender differences in upper secondary education cannot explain the differences in SweSAT scores. Within all the educational groups the male test takers outperform the females. However, the gender differences in SweSAT scores differ between the educational groups, being smaller among those who have entered a natural science or a technical programme. For the 1991 test there is also a difference between the other two educational groups in so far as the gender differences are greatest among those with a 2 year upper secondary programme.

Moreover, Table V also shows that keeping Lang under control leads to a substantial increase in the gender differences in SweSAT scores within the educational groups, and this is the same trend as was found for the total group. On the other hand, keeping Testsum under control does not seem to mean much as the gender differences are then of about the same magnitude as those when no control variable is introduced. The reason for this is to be found in Table III, where we can see that the gender differences in Testsum are small within the educational groups.

Thus, the conclusion to be drawn is that the male superiority in the SweSAT scores cannot be explained neither by gender differences in Lang, Testsum nor educational background. On the contrary, taking the marks in Swedish and English as predictors, the adjusted gender differences in favour of the male test takers even exceed the actual ones.

WORD

As was the case for the SweSAT, the marks in Swedish and English are those marks which have the highest predictive power for the WORD scores. Among the tests Op proved to be a more effective control variable than the Testsum, which is to be expected since both Op and WORD are tests of word knowledge.

Table VI. Gender differences in the WORD subtest.

Control variable	1991		1992	
	ES	R^2	ES	R^2
None	.15*	*.01*	.22*	*.01*
Lang	.36*	*.25*	.46*	*.27*
Op	.16*	*.38*	.26*	*.37*
Lang + Op	.26*	*.43*	.39*	*.43*

The gender differences in WORD are smaller than those in the total SweSAT scores, but are still statistically significant. As can be seen from the R^2-values, gender explains only 1 per cent of the variance in WORD as compared to 9 and 5 per cent respectively in the total SweSAT scores. The R^2-values also show that Op explains a substantial part of the variance in

WORD, but it does not change the gender differences more than marginally. Lang, on the other hand, is less predictive for the WORD scores, but it changes the gender differences substantially more. As a matter of fact, now the differences are more than twice as high as the actual ones, and this is because the females have the highest marks. Controlling for both Op and Lang simultaneously results in a gender difference between those obtained when the control variables are introduced separately.

Table VII. Gender differences in the WORD subtest by educational group.

Control variable	1991 NT	1991 HSE	1991 2YEAR	1991 R^2	1992 NT	1992 HSE	1992 2YEAR	1992 R^2
None	.03	.05	.18	*.05*	.12	.21	.30	*.10*
Lang	.33	.32	.42	*.26*	.42	.42	.48	*.29*
Op	.14	.05	.15	*.39*	.19	.29	.13	*.40*
Lang +Op	.25	.18	.27	*.43*	.34	.39	.27	*.45*

Taking previous education into account gives a somewhat different picture of the gender differences in WORD. For the 1991 test the gender differences nearly disappear among test takers with a long and theoretical upper secondary education (NT and HSE), while there still remains a difference in favour of male test takers among those with the same educational background who took the test 1992. On both test occasions the 2YEAR group shows a clear difference in favour of male test takers. Just as was the case for the total groups, controlling for Lang leads to a substantial increase of the gender differences also within the various educational categories. When this variable is controlled for, the male test takers outperform the females by 0.3 to 0.5 units of a standard deviation which correspond to about 1.5 and 2 raw score points, respectively.

Taking Op into consideration reduces the gender difference in WORD scores among those from the 2 year educational programmes who took the SweSAT in 1992. For the other subgroups the differences are either increased or remain unaffected. When both Op and Lang are controlled for the gender differences in WORD fall between those obtained when only one of these variables are involved.

Thus, males obtain the highest scores on WORD within the total group, but in a few instances these differences are small when educational background and Op is taken into account. If, on the other hand, the gender differences in Lang are accounted for, the male test takers' mean score substantially exceeds that for the females irrespective of educational background.

READ

When we now turn to the second verbal subtest, READ, the same control variables are included in the analyses as for WORD, that is to say Op and Lang.

Table VIII. Gender differences in the READ subtest.

Control variable	1991 ES	1991 R^2	1992 ES	1992 R^2
None	.29*	*.02*	.23*	*.01*
Lang	.49*	*.22*	.44*	*.22*
Op	.30*	*.30*	.27*	*.31*
Lang +Op	.40*	*.35*	.37*	*.35*

The actual gender differences in the READ subtest amount to 0.3 and 0.2 units of a standard deviation, respectively, and both these differences are statistically significant. Expressed as raw score points these differences are 1 and 0.7 points, respectively. As shown by the R^2-values gender explains only one or two per cent of the variance in READ scores. Adding Lang implies an increase in the proportion of explained variance by 20 per cent and adding Op implies an even greater increase. This is in agreement with the results found for WORD. As for WORD, taking Op into account does not change the gender differences much, but this is the case when the differences in Lang are controlled for. Now the male test takers outperform the female test takers by 0.5 and 0.4 units of a standard deviation. That is to say that, just as we found for WORD, according to Lang we should expect a difference in the READ scores in favour of the females, but the actual differences are in the opposite direction.

Table IX. Gender differences in the READ subtest by educational group.

Control variable	1991 NT	1991 HSE	1991 2YEAR	1991 R^2	1992 NT	1992 HSE	1992 2YEAR	1992 R^2
None	.11	.21	.31	*.08*	.00	.28	.54	*.09*
Lang	.37	.44	.51	*.23*	.27	.45	.71	*.23*
Op	.20	.21	.27	*.33*	.07	.34	.40	*.33*
Lang +Op	.31	.33	.39	*.36*	.19	.43	.51	*.37*

On the 1991 test the gender differences in favour of males remain even if they are relatively small for the NT group. On the 1992 test this group does not show any difference at all in the actual scores. When the control variables are introduced, however, all differences are in the favour of the males, and they are substantial when Lang constitutes the control variable. This means that the small or non-existent gender differences within the NT group are due to the fact that these females have a higher verbal ability as measured by Op and the language marks.

The results in Table IX also show a general trend of increasing gender differences when we go from NT over HSE to the 2YEAR group. Within the last mentioned group, for instance, the gender difference in the 1992 test in favour of males is not less than 0.7 units greater than what could be expected from the differences in Lang.

In sum then, the male test takers outperform the females also on the READ test and this difference cannot be fully explained by gender differences in marks or test scores. The same is true also for the various educational subgroups with only one exception - the NT group who took the test in 1992.

STECH

As mentioned in the introduction, STECH was replaced by the English reading comprehension test (ERC) in 1992. Therefore, the gender differences in STECH can only be studied for the first test occasion.

Table X. Gender differences in the STECH subtest.

Control variable	Total sample		Educational groups			
	ES	R^2	NT	HSE	2YEAR	R^2
None	.41*	*.04*	.34	.26	.48	*.07*
Lang	.59*	*.23*	.59	.50	.69	*.23*
Testsum	.30*	*.23*	.33	.24	.47	*.24*
Lang +Testsum	.46*	*.30*	.51	.42	.62	*.30*

The actual gender difference in the STECH subtest amounts to 0.4 units of a standard deviation which corresponds to nearly 1.5 raw score points. To some extent this difference is explained by the gender difference in Testsum, but even when this variable is controlled for the males achieve significantly higher scores than the females. When the marks and the marks in combination with Testsum are kept under control the adjusted gender difference becomes greater than the actual one. Thus, there are substantial gender differences in favour of the male test takers in the total sample also on STECH and these differences cannot be explained by differences in marks or Testsum.

The same conclusion is valid also for each of the educational groups. The smallest difference is found for the HSE group, where the actual difference amounts to a quarter of a standard deviation unit and the greatest differences of about half a standard deviation unit is shown by the 2YEAR group. Controlling for test scores does not mean much for the differences and controlling for marks implies increased differences.

The R^2 values show that controlling for upper secondary education does not imply any greater increase in the proportion of explained variance, and the small increase found is almost wholly due to differences in test results and marks. When these variables are controlled for the proportions of explained variance within the educational groups equal those for the total group.

ERC

Since ERC replaced the STECH in the 1992 test, it is of great interest to make a comparison between these two subtests concerning the gender differences. Did this change of content in the SweSAT imply any changes in gender differences in the total test score?

Table XI. Gender differences in the ERC subtest.

Control variable	Total sample		Educational groups			
	ES	R^2	NT	HSE	2YEAR	R^2
None	.37*	*.03*	.11	.42	.55	*.13*
Lang	.61*	*.28*	.41	.63	.73	*.31*
Op	.40*	*.26*	.17	.48	.41	*.31*
Op + Lang	.56*	*.36*	.35	.61	.58	*.38*

The introduction of ERC instead of STECH seems not to have changed the gender differences more than marginally within the total sample. The males outperform the females by about 0.4 units of a standard deviation also in ERC, and this difference cannot be explained neither by gender differences in the marks nor in the Op scores. On the contrary, keeping marks under control leads to an increase in the ERC scores to 0.6 units of a standard deviation. Thus, the male superiority in the ERC scores becomes even more pronounced when we take the gender differences in Swedish and English into account.

Even if ERC has not changed the differences within the total sample, it has had some consequences for those gender differences found within the various educational groups. Among those who have finished an NT programme the gender differences are smaller in ERC than in STECH. This is particularly the case for the actual difference which has decreased from 0.3 units of a standard deviation in STECH to 0.1 in ERC. On the other hand, the differences have increased somewhat within the other two educational groups, and a comparison between all three groups shows that the gender differences throughout are smallest within the NT group.

In the same way as for the total sample, the gender differences within the educational groups cannot be explained by differences in test scores or marks. As shown by table XI the differences are from 0.4 to 0.7 units greater than what could be expected from the marks in English and Swedish and they are from 0.2 to 0.5 units greater than what the differences in Op give occasion to expect. Controlling for Op and Lang simultaneously certainly implies an increase in the proportion of explained variance in ERC, but at same time the gender differences become greater than when keeping only Op under control.

GI

Even if GI is a test of general information and not primarily a verbal test the marks in Swedish and English together with Opposites are the control variables which have shown the best predictive power. Therefore, these variables are included in the analyses of GI.

Table XII. Gender differences in the GI subtest.

Control variable	1991 ES	1991 R^2	1992 ES	1992 R^2
None	.57*	*.07*	.27*	*.02*
Lang	.74*	*.23*	.45*	*.17*
Op	.57*	*.28*	.30*	*.21*
Lang +Op	.67*	*.31*	.40*	*.25*

The actual gender differences in the GI scores vary substantially between the two test occasions. On the 1991 test they amount to more than half a standard deviation unit in favour of male test takers, but on the 1992 test they are only half as great. However, even on the latest test occasion they are statistically significant. Expressed as raw score points these gender differences are 2 and 1 point, respectively. Keeping Lang under control implies an increase in the differences for both the tests occasions which is natural in the light of the females' higher marks in Swedish and English. Controlling for Op, on the other hand, implies practically no change.

The rather great discrepancy in gender differences between the two test occasions has no clear-cut explanation. One possible cause may be the composition of the groups, but this explanation is hardly supported by the gender differences found for the WORD and the READ subtests, nor does Table III indicate smaller gender differences among those who took the SweSAT in 1992. Therefore, the composition of the groups seems not to be the most likely explanation. Another possible cause may be the test content. GI is a test which covers a wide content area and what areas that happen to be represented in a specific test may well have consequences for the gender differences. In an analysis of "male and female items" for one test occasion Stage (1996, p 12) found that GI together with WORD were the subtests which showed the greatest variation of items in the female-male dimension.

Table XIII. Gender differences in the GI subtest by educational group.

Control variable	1991 NT	1991 HSE	1991 2YEAR	1991 R^2	1992 NT	1992 HSE	1992 2YEAR	1992 R^2
None	.36	.51	.57	*.12*	.18	.22	.14	*.11*
Lang	.59	.73	.75	*.24*	.40	.37	.27	*.13*
Op	.43	.51	.54	*.30*	.23	.28	.02	*.26*
Lang +Op	.54	.63	.65	*.33*	.34	.35	.11	*.28*

The actual gender differences found for the total sample on the 1991 test are valid also within each educational group, even if they tend to be somewhat smaller for those who have finished the NT programmes. For the 1992 test, on the other hand, the gender differences within all educational groups are lower than that for the total group and for those who have finished a 2-year programme the difference is almost completely explained by the gender differences in Op. For all other groups Op does not contribute at all to the explanation of the gender differences. Just as was the case for the total sample, keeping Lang under control leads to an increase of the gender differences in all subgroups.

Thus, the conclusion to be drawn concerning the GI subtest is that there are clear gender differences in favour of males - differences that cannot be explained by gender differences in marks and ability except in one case, namely among those who have finished a 2-year programme of upper secondary school and who took the test in 1992. For them the differences in GI scores are almost completely explained by the gender differences in Op.

DS

The two subtests which remain to be analysed are DS and DTM, that is to say the two quantitative subtests. For both these subtests Testsum and the marks in the natural science subjects (Mathsci) will constitute the control variables.

Table XIV. Gender differences in the DS subtest.

Control variable	1991		1992	
	ES	R^2	ES	R^2
None	.78*	*.11*	.67*	*.08*
Mathsci	.64*	*.35*	.62*	*.17*
Testsum	.65*	*.34*	.59*	*.35*
Mathsci + Testsum	.65*	*.36*	.58*	*.36*

In comparison to the previously studied subtests and also in comparison to the total SweSAT score DS shows the greatest gender differences on both test occasions. For the 1991 test the males outperform the females by nearly 0.8 units of a standard deviation. For the later test occasion the difference is one tenth of a unit smaller. These differences correspond to 2.5 and 2.1 raw score points, respectively. Controlling for the marks in the natural science subjects and for Testsum reduces the difference somewhat, but even when account is taken of the gender differences in these variables, the difference in DS scores is about 0.6 units of a standard deviation.

Table XV. Gender differences in the DS subtest by educational group.

Control variable	1991				1992			
	NT	HSE	2YEAR	R^2	NT	HSE	2YEAR	R^2
None	.39	.63	.77	*.24*	.35	.60	.53	*.25*
Mathsci	.42	.71	.64	*.38*	.38	.57	.54	*.28*
Testsum	.39	.61	.76	*.40*	.37	.57	.45	*.42*
Mathsci + Testsum	.39	.61	.77	*.40*	.38	.56	.46	*.42*

The gender differences in the total samples are also found within each of the educational groups. However, the differences are normally smaller and particularly so for the NT group, where they amount to about 0.4 units of a standard deviation. As shown by the uniform coefficients in Table XV, controlling for gender differences in Mathsci and Testsum makes no or only very small contributions to the explanation of the differences.

Worth noticing is the substantial increase in the R^2 values from Table XIV to Table XV when no control variable is included. This increase seems to indicate that the programme chosen in upper secondary school is a factor of some importance to the DS scores. However, to a large extent this influence occurs via the marks and Testsum. Therefore, just as was the case for the total SweSAT score, the increase in R^2 values seems to a large extent to originate from those differential selection processes which occur in the transition from compulsory school into the various programmes of upper secondary school.

Thus, we can conclude that there are very great gender differences in DS scores and that they cannot be explained by gender differences in the control variables or in the test takers' previous education.

DTM

As mentioned in the introduction, DTM is the subtest which besides DS has always shown the greatest gender differences in favour of male test takers. This is also the case for the present two test occasions. As shown by Table XVI the gender difference for the 1991 test even exceeds that for the DS subtest, while it is somewhat lower for the 1992 test. Expressed as raw score points the differences amount to 2.5 and 1.8 points, respectively.

Table XVI. Gender differences in the DTM subtest.

Control variable	1991		1992	
	ES	R^2	ES	R^2
None	.85*	*.15*	.62*	*.07*
Mathsci	.82*	*.20*	.57*	*.15*
Testsum	.72*	*.40*	.55*	*.32*
Mathsci + Testsum	.71*	*.41*	.53*	*.34*

Controlling for Testsum and marks, separately or simultaneously, leads to a small decrease in gender differences, but still the male test takers perform significantly higher than do the female test takers.

According to the R^2 values the marks contribute moderately to the explanation of the variance in DTM. Testsum, on the other hand, makes a substantially greater contribution, and this may be caused by the fact that DTM does not primarily measure traditional school knowledge and the strategies required for item solutions are of a rather general nature.

From Table XVI we can also see the same trend as for GI, namely that the gender differences are smaller in the 1992 test as compared to those in the test given in 1991. Just as was the case for GI the explanation is not clear, but the most probable cause is that the content of the DTM subtest differs between the two test occasions in such a way that it has influenced the gender differences. Like GI, DTM is a test which covers a wide content area and it seems likely that variations in content may influence the gender differences.

Table XVII. Gender differences in the DTM subtest by educational group.

Control variable	1991				1992			
	NT	HSE	2YEAR	R^2	NT	HSE	2YEAR	R^2
None	.62	.79	.85	*.15*	.32	.66	.23	*.25*
Mathsci	.64	.80	.88	*.22*	.34	.63	.25	*.27*
Testsum	.62	.77	.83	*.41*	.34	.63	.17	*.39*
Mathsci + Testsum	.62	.77	.85	*.41*	.34	.62	.17	*.40*

Taking previous education into consideration leads to smaller gender differences in the DTM scores among those who have finished the NT programmes and particularly so for the 1992 test. On this occasion, however, the gender differences are even smaller within the 2YEAR group, while they remain of the same magnitude in the HSE group as for the total sample. On the whole, controlling for marks and Testsum leads to only small changes.

Consequently we can conclude that the great gender differences in DTM scores are only to a small extent explained by differences in test scores and marks from compulsory school. Previous education seems to play a somewhat more important role in so far as the gender differences are smaller for the NT group than for the total sample. The same is true also for the 2YEAR group in the 1992 test. However, as we have found previously, it seems not to be the knowledge acquired in upper secondary school that is of importance, but rather the selection process to this educational level.

DISCUSSION

Bias has often been put forward as an explanation of the fact that the male SweSAT takers have always achieved higher scores than have the females. However, mean differences between self selected groups like the SweSAT takers can never be taken as indicators of bias in the test unless relevant corrections have been made. In this study such corrections are based on the leaving marks from compulsory school and on test scores from the age of 13. Furthermore, upper secondary education is taken into consideration by comparing the adjusted SweSAT scores for test takers within the various upper secondary programmes.

The main results for the two test occasions are summarised in Figures 1 and 2.

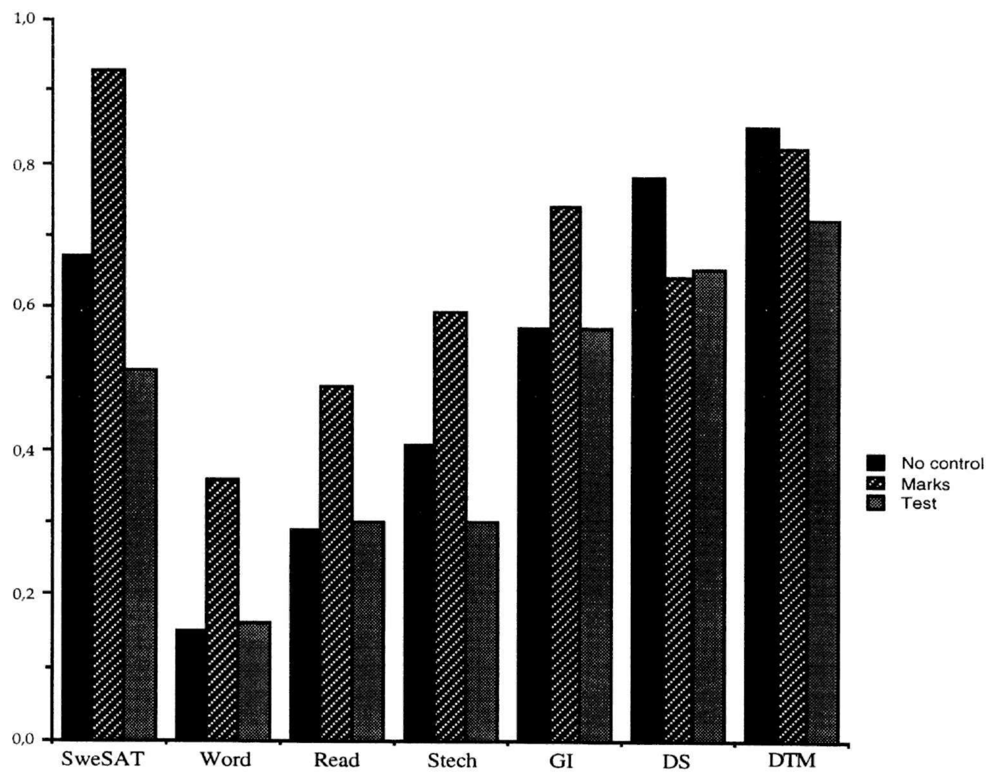

Figure 1. Actual and adjusted gender differences in SweSAT scores on the 1991 test. Effect sizes.

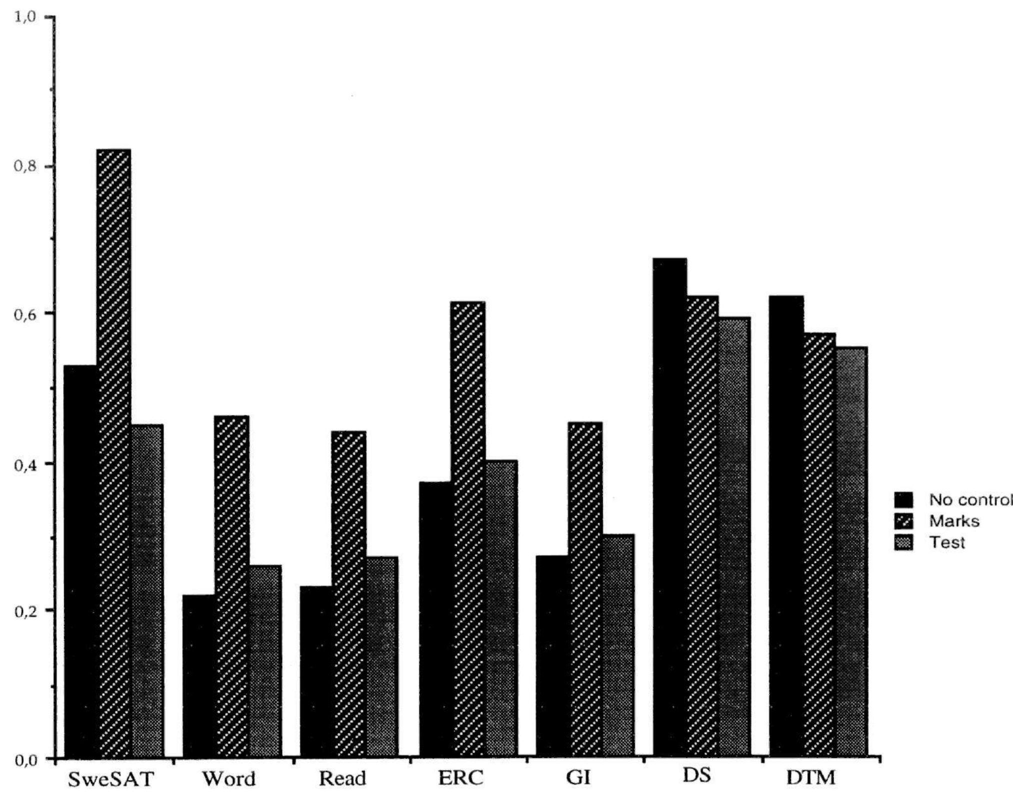

Figure 2. Actual and adjusted gender differences in SweSAT scores on the 1992 test. Effect sizes.

As both figures show, there are substantial gender differences in favour of the male test takers in the total score (SweSAT), and these differences cannot be explained by differences in marks or differences in test scores. When the marks are kept under control the adjusted differences even exceed the actual ones, which is due to the fact that the females normally receive the highest marks in the compulsory school. Taking test scores into account implies somewhat smaller gender differences, but the male test takers still score substantially higher than do the female test takers.

Looking at the separate subtests the picture remains much the same. For the 1991 test both the actual and the adjusted gender differences increase regularly when we go from WORD to the quantitative subtests DS and

DTM. On the later test occasion this trend is not equally regular due to the fact that the GI subtest now shows a rather small gender difference. On both test occasions, however, the gender differences are greater on the quantitative parts than on the verbal parts, and this result is in agreement with what has been found in other studies (Hyde, Fennema & Lamon, 1990; Cleary, 1992; Stage, 1992).

Can the gender differences in SweSAT scores be explained by the fact that males and females have taken different programmes in upper secondary school? The answer to that question is no. Certainly, we have found a few instances of small actual gender differences on the verbal subtests and few such instances also when test scores have been taken into consideration. Only once has a gender difference of 0.00 been found, however, and that is the actual difference on the READ subtest in spring 1992 among test takers from the NT upper secondary programmes. This instance of equal means for female and male test takers, however, should be seen in the light of females having higher verbal test scores and higher language marks from compulsory school. When these variables are controlled for the differences in READ scores are in favour of the male test takers.

There is also a general trend of the smallest gender differences within the NT group. In this respect the results differ from those found by Stage (1988). According to her results the gender differences were smallest among students of humanities and greatest among students of economics. The discrepancy in results are, of course, due to the fact that in the present study these two categories have been brought together into the same category. One possible explanation for the comparatively small gender differences within the NT group is the differential selection effects in the transition from compulsory school to upper secondary education.

On the whole, the upper secondary school programme seems to be of little importance to SweSAT performance. Certainly, the proportion of explained variance is increased by taking upper secondary education into account, but this effect fades out when the test scores or the marks are also involved. This fact indicates that the effect of upper secondary education is not to be ascribed to the knowledge acquired on this educational level, but instead to the selection processes which are at work when the individual chooses upper secondary education. This might seem somewhat astonishing, but as a matter fact, the SweSAT is not intended to measure that kind of school knowledge, but rather knowledge acquired in compulsory school.

The general conclusion to be drawn then, is that the gender differences in the SweSAT scores cannot be explained by gender differences in upper secondary education or by differences in test scores or marks from compulsory school. Does this mean that there is a bias in the SweSAT in favour of male test takers?

There is no doubt that the results from the present study support such a conclusion. However, it is important to state that these results do not prove the existence of a general bias in the SweSAT in favour of male test takers. There are several issues to be taken into consideration before any definite conclusions are drawn concerning bias. For instance:

- The validity of the results are highly dependent on the validity of the control variables.

- The groups studied are homogeneous as to age and previous education.

- Only two test occasions have been studied.

As to the validity of the control variables the results have shown that together with gender the control variables explain more than 50 per cent of the variance in the total SweSAT score and about 40 per cent of the variance in most of the subtests. Taking into consideration that no variable has perfect reliability and that the group studied constitutes a positively selected sample, these proportions must be regarded as high. Thus, the control variables used seem to be highly relevant and other even more relevant variables are not easily found, all the more as they must not be influenced by factors related to SweSAT taking i.e. differential educational experiences. Thus, these facts indicate a good validity in the results.

The group studied is homogeneous in age and it is not representative for all SweSAT takers. All test takers studied are in the age of 19 or 20 and as shown by Stage (1992) the gender differences seem to be larger among younger and older test takers than they are in the age span of 25-39 years. However, even if the gender differences in SweSAT scores vary with age, Stage and Jarl (1996) have shown that nearly 50 per cent of the test takers in spring 1994, 1995 and 1996 are in the ages up to 20 and that 25 to 30 per cent of them are 21 to 24 years old. Thus, there is no doubt that the results in the present study may be regarded as valid for a majority of SweSAT takers.

The group is homogeneous also with respect to educational background. All persons in this study have finished a 2- or 3-year programme of upper secondary school, and according to Stage's results this category of test takers show greater gender differences than do test takers without any upper secondary education, and smaller differences than test takers with some higher education. Stage and Jarl (1996), however, have shown that not less than 80 per cent of the test takers have finished a 2- or 3-year upper secondary programme. As to the validity of the results the same argument as that for age groups is applicable This argument, however, does not exclude the need for replicating the study on the basis of more heterogeneous samples, and according to our intentions this will be done in a near future.

Another objection which may be raised is that only two test occasions have been studied and both are the spring administrations of the SweSAT. It is quite possible that the groups who take the test in the autumn are somewhat different from those who take the test in spring. At least we know that there are substantial differences in the number of testees. Moreover, this study has shown that the gender differences are not perfectly stable from the first to the second test occasion, and the most probable explanation is variation in test content. Therefore, it seems reasonable to analyse more than two test occasions before any secure generalisations are made. On the other hand, the results have been uniform in so far as all gender differences have been in favour of the male test takers and are mostly considerable. Furthermore, Stage (1992) and Ingerskog and Stage (1993) have shown that there always has been a difference in favour of the male test takers. Therefore, there are no reasons to believe that the results found in this study would be unrepresentative for the gender differences in SweSAT scores. However, according to our intentions, more test occasions will be studied in a similar way when the data now available have been supplemented.

Finally, all variables studied are manifest variables and this fact implies two problems. One problem was mentioned before and it is the measurement errors which decrease the correlations between the control variables and the SweSAT scores with the ensuing undercorrections for initial differences. Another problem connected to the manifest variables is the fact that the results are only descriptive in nature. We know that the females have underachieved in relation to what could be expected from their test scores and marks from compulsory school, but the results say nothing about the causes. Are the gender differences in SweSAT scores due to differences in the main factors measured by the test or are there specific elements in the test which cause these differences? In a forthcoming study we are going to analyse the latent variables and then we will be able to avoid these two problems.

REFERENCES

Bränberg, K., Henriksson, W., Nyquist, H., & Wedman, I. (1990). The influence of sex, education and age on test scores on the Swedish Scholastic Aptitude Test. *Scandinavian Journal of Educational Research, Vol. 34,*No. 3, 189-203.

Cleary, A. (1992). Gender differences in aptitude and achievement test scores. In J. Pfeiderer (Ed.), *Sex equity in educational opportunity, achievement, and testing. Proceedings of the 1991 ETS invitational conference.* (pp 51-90). Princeton, NJ: Educational Testing Service.

Cole, N. S. (1997). *The ETS gender study: How females and males perform in educational settings.* Princeton, NJ: Educational Testing Service.

Cole, N. S., & Moss, P. A. (1993). Bias in test use. In R. L. Linn (Ed.), *Educational measurement 3 rd. ed.* (pp. 201-219). Phoenix, AZ: Oryx Press.

Emanuelsson, I., Reuterberg, S-E., & Svensson, A. (1993). Changing differences in intelligence? Comparisons between groups of 13-year-olds tested from 1960 to 1990. *Scandinavian Journal of Educational Research, Vol. 37,*No. 4, 259-277.

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist, Vol. 43,*No. 2, 95-103.

Henriksson, W., Stage, C., & Lexelius, A. (1986). Samma uppgift men olika innehåll - en studie av NOG-provet. (The same item but a different content - a study of the Data Sufficiency subtest). *Pedagogiska institutionen, Umeå universitet, PM 10.*

Howe, H. (1985). Let's have another SAT score decline. *Phi Delta Kappan, 66,* 599-602.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107,* 139-155.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104,* 53-69.

Härnqvist, K. (1993). *Den sociala selektionen till gymnasieskola och högskola. Analyser i den longitudinella databasen UGU 67.* [The social selection to upper secondary school and to higher education. Analyses in the longitudinal data base ETF 67] (Rapport 1993:10). Göteborg: Göteborgs universitet, Institutionen för pedagogik.

Härnqvist, K., Emanuelsson, I., Reuterberg, S-E., & Svensson, A. (1994). *Dokumentation av projektet "Utvärdering genom uppföljning"* [Documentation of the project "Evaluation Through Follow-up"] (Rapport 1994:3). Göteborg: Göteborgs universitet, Institutionen för pedagogik.

Ingerskog, P., & Stage, C. (1993). Högskoleprovet hösten 1993. Provtagargruppens sammansättning och resultat. (The Swedish Scholastic Aptitude Test in the autumn of 1993. The characteristics and the results of the the group of test takers). *Pedagogiska institutionen, Umeå universitet, PM 7.*

Linn, M. C. (1991). Gender differences in educational achievement. In J. Pfeiderer (Ed.), *Sex equity in educational opportunity, achievement, and testing. Proceedings of the 1991 ETS invitational conference* (pp. 11-50). Princeton, NJ: Educational Testng Service.

Maccoby, E., & Jacklin, C. (1974). *The psychology of sex differences.* Stanford: Stanford University Press.

Mäkitalo, Å. (1994). *Non-comparability of female and male admission test takers* (Rapport No. 1994:06). Göteborg, Sweden: Göteborg University, Department of Education and Educational Research.

Mäkitalo, Å. (1996). Gender differences in performance on the DTM subtest in the Swedish Scholastic Aptitude Test as a function of item position and cognitive demands. *Scandinavian Journal of Educational Research, Vol. 40,* No. 3, 189-201.

Mäkitalo, Å., & Reuterberg, S-E. (1996). *Who takes the Swedish Scholastic Aptitude Test? A study of differential selection to the SweSAT in relation to gender and ability* (Rapport No. 1996:03). Göteborg, Sweden: Göteborg University, Department of Education and Educational Research.

Reuterberg, S-E. (1994). *Selection of Swedish Scholastic Aptitude Test takers as a function of social background and ability* (Rapport No. 1994:05). Göteborg, Sweden: Göteborg University, Department of Education and Educational Research.

Reuterberg, S-E. (1996). Selektionsmekanismernas betydelse för gruppskillnader på Högskoleprovet. (The importance of selection mechanisms to group differences on the Swedish Scholastic Aptitude Test.) In *The Swedish Scholastic Aptitude Test. Through the eyes of eleven researchers.* National Agency for Higher Education, 1996:22R.

Rosenthal, R., & Rubin, D. B. (1982). Further meta-analytic procedures for assessing cognitive gender differences. *Journal of Educational Psychology, Vol. 74,* No. 5, 708-712.

Stage, C. (1987). Analys av NOG-uppgifter med avseende på könsskillnader i resultat. (Analysis of Data Sufficiency items with respect to gender differences in scores). *Pedagogiska institutionen, Umeå universitet, PM 12.*

Stage, C. (1992). How important are age and education for gender differences in test results? *Scandinavian Journal of Educational Research, Vol. 36,* No. 3.

Stage, C. (1996). Samma poäng -samma prestation? En jämförelse på delprovs- och uppgiftsnivå mellan män och kvinnor med samma resultat på högskoleprovet. (Same score - same achievement? A comparison on subtest and item level between men and women with the same score on the SweSAT.). *Enheten för pedagogiska mätningar, Umeå universitet, PM 113.*

Stage, C., & Jarl, C. (1996). Högskoleprovet våren 1996. Provdeltagarnas sammansättning och resultat. (The SweSAT in the spring of 1996. The characteristics and the results of the group of test takers). *Enheten för pedagogiska mätningar, Umeå universitet, PM 115.*

Stumpf, H., & Stanley, J. C. (1996). Gender-related differences on the College Bord's advanced placement and achievement tests, 1982-1992. *Journal of Educational Psychology, Vol. 88,*No. 2, 353-364.

Svensson, A. (1971). *Relative achievement. School performance in relation to intelligence, sex and home environment.* Stockholm: Almqvist & Wiksell.

Wainer, H. (1986a). The SAT as a social indicator: A pretty bad idea. In H. Wainer (Ed.), *Drawing inferences from self-selected samples* (pp. 7-21). New York: Springer Verlag.

Wainer, H. (1986b). Five pitfalls encountered while trying to compare states on their SAT scores. *Journal of Educational Measurement, Vol. 23, 22-24.*

Wainer, H. (1993). Does spending money on education help? A reaction to the Heritage Foundation and the Wall Street Journal. *Educational Researcher, 9,* 22-24.

Wainer, H., & Steinberg, L. S. (1992). Sex differences in performance on the mathematics section of the Scholastic Aptitude Test: A bidirectional validity study. *Harvard Educational Review, Vol. 62,*No. 3, 323-336.

Wedman, I. (1994). The Swedish Scholastic Aptitude Test: Develpoment, use, and research. *Educational Measurement: Issues and practice. Vol. 13,*No. 2, 5-11.

Wester-Wedman, A. (1992a). Item bias with respect to gender interpreted in the light of problem-solving strategies. *Department of Education, University of Umeå, EM 3.*

Wester-Wedman, A. (1992b). Ett försök med öppna frågor i DTK-provet. En jämförelse mellan öppna frågor och flervalsfrågor avseende könsskillnader i prestation på DTK-provet. (An experiment with open-ended questions in the DTM subtest. A comparison between open-ended questions and multiple choice questions with respect to the gender difference in performance on the DTM subtest). *Pedagogiska institutionen, Umeå universitet, PM 54.*

Wester-Wedman, A. (1992c). Förlängd provtid på DTK-provet. En studie av effekten av förlängd provtid på könsskillnaden i prestation på DTK-provet. (Prolonged testing time on the DTM subtest. A study of the impact of prolonged testing time on the gender difference in performance on the DTM subtest). *Pedagogiska institutionen, Umeå universitet, PM 56.*

Wilder, G., & Powell, K. (1989). Sex differences in test performance: A survey of the literature. *College Board Report, No. 89-3.*

APPENDIX

Table AI. Correlations between SweSAT scores and testscores/marks. SweSAT -91.

	Op	Ns	Testsum	Lang	Mathsci	Sach
SweSAT	.61	.49	.64	.48	.31	.32
WORD	.61	.25	.47	.47	.17	.22
READ	.53	.32	.47	.41	.24	.28
STECH	.46	.36	.46	.39	.23	.22
GI	.45	.29	.43	.33	.24	.24
DS	.31	.48	.51	.24	.28	.25
DTM	.36	.51	.55	.28	.25	.22

Bold figures show highest coefficients.

Table AII. Correlations between SweSAT scores and testscores/marks. SweSAT -92.

	Op	Ns	Testsum	Lang	Mathsci	Sach
SweSAT	.61	.43	.60	.50	.34	.34
WORD	.60	.25	.44	.47	.23	.26
READ	.54	.25	.42	.42	.23	.26
ERC	.48	.27	.41	.46	.25	.27
GI	.44	.26	.41	.35	.22	.24
DS	.34	.48	.53	.25	.31	.27
DTM	.31	.51	.52	.28	.30	.27

Table AIII. Proportion of explained variance (R^2) for different combination of control variables.

	Test/mark-combination	R^2 SwSAT-91	R^2 SwSAT-92
SweSAT	Op + Lang	.42	.42
	Testsum + Lang	**.46**	**.43**
Word	**Op + Lang**	**.41**	**.40**
READ	**Op + Lang**	**.31**	**.32**
STECH	Op + Lang	.24	
	Testsum + Lang	**.25**	
ERC	**Op + Lang**		**.30**
GI	**Op + Lang**	**.22**	**.22**
	Testsum + Lang	.21	.21
DS	**Testsum + Mathsci**	**.28**	**.30**
DTM	Testsum + Lang	.30	.27
	Ns + Lang	.28	.28
	Testsum + Mathsci	**.31**	**.28**
	Ns + Mathsci	.27	.28

Bold=the combination chosen as control

REPORTS FROM DEPARTMENT OF EDUCATION AND EDUCATIONAL RESEARCH, GÖTEBORG UNIVERSITY

ISSN 0282-2156

To be ordered from Department of Education and Educational Research, Göteborg University, P.O. Box 1010, S-431 26 MÖLNDAL, Sweden.

Asplund Carlsson, M., Halász, L., Marton, F. Some Paradoxical effects of attempts to facilitate readers' understanding of a Kafkaesque paradox. 1994:01.

Holmberg, L. M. The thoughts and practices of user documentation developers. 1994:02.

Holmberg, L. M. A content analysis of nine user documentations on operating systems. 1994:03.

Ahlberg, A. & Csocsán, E. Grasping numerosity among blind children. 1994:04.

Reuterberg, S-E. Selection of Swedish Scholastic Aptitude Test takers as a function of socioeconomic background and ability. 1994:05.

Mäkitalo, Åsa. Non-comparability of female and male admission test takers. 1994:06.

Rosén, M. Gender differences in hierarchically ordered ability dimensions: the impact of missing data. 1995:01.

Holmberg, L.M. The design of a learning environment for users of an accounting information system. Theoretical foundations and an empirical study. 1996:01.

Holmberg, L.M. Design of learning environments for IT-users. 1996:02.

Mäkitalo, Å. & Reuterberg, S-E. Who takes the Swedish scholastic aptitude test? A study of differential selection to the SweSAT in relation to gender and ability. 1996:03.

Hägglund, S., Hakvoort, I. & Oppenheimer, L. (Eds.) Research on children and peace: international perspectives. 1996:04.

Hasselgren, B. & Beach, D. Phenomenography: A "good-for-nothing brother" of phenomenology? or: Phenomenography is what phenomenographers do when doing phenomenography. 1996:05.

Folkestad, G., Lindström, B. Hargreaves, D.J. Colley, A. & Comber, C. Gender, computers and music technology. Experience and Attitudes. 1996:06.

Alexandersson, C. & Beach, D. Students´ experiences of different pedagogical practices. - A study of Personnel Manager Programme: Pedagogics recruitment fall off and course quality issues. 1997:01.

Reuterberg, S-E. Gender differences on the Swedish scholastic aptitude test. 1997:02.

Digitaltryckt & Bunden
Vasastadens Bokbinderi AB
1997