Gender bias in The Swedish Scholastic Assessment Test for entrance to university?

Essay in Statistics, Spring 2007

Department of Economics and Society, Dalarna University

Author **Boyuan Zhao**

Supervisor **Johan Bring**

Date **June, 2007**

Abstract

The Swedish Scholastic Assessment Test, SweSAT, is an admission test used for selection to higher education. After looking at the result of the last 10 test-occasions, we can find that men have always (on average) performed better than women. This fact has lead to a lot of debate and controversies. In this paper, there are some statistical methods used to test if there is gender bias in SweSAT.

The statistical methods used were descriptive statistics, t-test and confidence intervals. The results showed that a) there is gender bias in SweSAT, b) there are four of the five subtests that have a bias, c) there are no specific questions causing the gender bias but rather a general tendency among all questions within a subtest.

Key words:

Gender bias, SweSAT, Descriptive statistics, Normalized mean, Shapiro-Wilk test, T-test, Interval estimation, Confidence interval.

1. Introduction

1.1. Background

"The Swedish Scholastic Assessment Test, SweSAT, is an admission test used for selection to higher education." (http://www.umu.se/edmeas/hprov/index_eng.html, 2007-5-30) Since it is a selection test, it is supposed to rank the applicants as fairly as possible according to their expected academic success. The content of the test does not reflect any specific curriculum, although it is designed to be consistent with school-based learning.

The SweSAT consists of 122 multiple-choice items distributed over five subtests. They are the comprehension of words and concepts (the subtest WORD), mathematical reasoning ability (DS, i.e. Data Sufficiency), Swedish reading comprehension (READ), the ability to interpret diagrams, tables, and maps (DTM), and English reading comprehension (ERC).

1.2. Problems raised

We have data on an individual level for the last ten test-occasions. Approximately 40 000 subjects take the test each time. Looking at the data, we can find that men have always (on average) performed better than women. This fact has lead to a lot of debate and controversies.

What I have done is analyze the data and find out if there is gender difference and point out the reason of the gender difference.

1.3. Aims

The aim of this paper is to describe and analyze the gender difference in the results. To get this aim, there are some questions to be solved:

a. Is there gender difference?

This question is to ask that if there is significant difference between the result of male and

the result of female.

b. Is there some specific sub-test?

Since there are five sub-tests in a test-occasion, if there is some specific sub-test has lead to gender difference of that test-occasion is wondered. On the other word, this question is to ask if there is significant difference between the result of male and the result of female in each sub-test.

c. Is there some specific questions?

Since male and female are interest in different fields, such as male always pay more attention on sports, while female always pay more attention on fashion. This question is to ask what the reason of gender difference, which means that if the gender bias of result are caused of some specific questions about specific fields, or do not have relationship with what the questions are talking about.

2. Data and methods

There are 10 data files which belong to 10 tests from spring 2002 to autumn 2006. The size of the database is 24 variables with 368402 cases.

2.1. Data import and organize

2.1.1. Missing data

In the first step, missing values were replaced by '999'.

Then, looking at the column of YEAR, there are 2 cases which have 11 as YEAR in the raw data (ID of the two cases are 263795 and 329984), which we can deal as missing values. So the two observations with year=11 were deleted since those values are unrealistic.

At last, look at the columns of the scores of five sub-tests, total, and normal.

There are 4 data files without missing data in columns of ORG, NOG, LAS, DTK, ELF, TOTAL, and NORMAL, which are data of spring 2002, spring 2004, autumn 2004, and spring 2006. Now, pay attention to the cases with missing value of the other 6 data files.

The reason of missing data is that the candidate has 0 correct answers. Here, it is not sure of the reasons of 0 score, such as the candidate had not take the test, or the candidate had took the test but can not give even one correct answer. The number of cases with missing data is no more than 200, and the total number of cases is 368400. It is only 0.054%. So those with zero correct answers since it's likely that they are missing observations are deleted.

2.1.2. Data import and organize

After dealing with missing data, the 10 files are put into one file. After name each variables, the data was imported in R. We can see the names of each variable in table 1(Appendix1).

In this paper, the aim is to analysis the difference of test results of male and female, the data of gender, scores of five subtests, the total score, and the normal points are most important.

2.1.3. Meaning of important variables

2.1.3.1. Gender

In this column, there are 165415 men and 202987 women.

2.1.3.2. ORD and ORDU

ORD is the subtest about vocabulary (WORD). ORD measures the comprehension of words and concepts. It consists of 40 items in which a word or phrase is given, and the task is to identify which of five options has the same, or almost the same, meaning. Words of both Swedish and foreign origin are included in this subtest. Range of ORD is (0, 40).

ORDU is the individual results for each question of ORD test, where “1” means right and “0” means wrong.

2.1.3.3. NOG and NOGU

NOG is the subtest about data sufficiency (DS). NOG aims at measuring mathematical reasoning ability. Each of the 22 items presents a problem and two statements. The task is to decide whether the statements provide enough information for solving the problem. Range of NOG is (0, 22).

NOGU is the individual results for each question of NOG test, where “1” means right and “0” means wrong.

2.1.3.4. LAS and LASU

LAS is the subtest about reading comprehension (READ). LAS measures Swedish reading comprehension in a broad sense. The test consists of five texts with four multiple-choice questions to each text, in total 20 items. The length of each text is roughly one page. Some questions concern details stated in the text, but most of them are designed to test the comprehension of larger parts or the text as a whole. Range of LAS is (0, 20).

LASU is the individual results for each question of LAS test, where “1” means right and “0” means wrong.

2.1.3.5. DTK and DTKU

DTK is the subtest about interpretation of diagrams, tables, and maps (DTM). DTK consists of ten sets of tables, graphs, or maps presenting information about different topics. There are two multiple-choice items to each set, which makes a total of 20 items. The degree of complexity varies from simply reading off a presented graph to problem-solving, i.e. processing information from all the different sources in the material. Range of DTK is (0, 20).

DTKU is the individual results for each question of DTK test, where “1” means right and

"0" means wrong.

2.1.3.6. ELF and ELFU

ELF is the subtest about English reading comprehension (ERC). ELF is of the same general type as the subtest READ. However, this subtest is more varied regarding both texts and item format. It consists of eight to ten texts which vary in length. Most of them are followed by one or more four-choice questions. The last text in the subtest has sentences where a word or a set of words has been omitted. The test taker is requested to choose the one of four options that best fits the rest of the sentence. The total number of items is 20. Range of ELF is (0, 20).

ELFU is the individual results for each question of ELF test, where "1" means right and "0" means wrong.

2.1.3.7. TOTAL

TOTAL is the total score of all the five subtests (ORD, NOG, LAS, DTK, and ELF). Range of TOTAL is (0, 122).

2.1.3.8. NORMAL

Since the degree of difficulty for each test-occasion is not the same, there is a normal score to get a standard score to make the results of different tests comparable. "When all the results have been obtained, the scores are standardized in order to equate results from different testing. Standardization means that the total number of correct answers is transfigured into a standard score ranging from 0.0 to 2.0, where the degree of difficulty to reach a certain score should always be the same, regardless of the variations in the population of testers or in the level of difficulty of the test."(http://www.umu.se/edmeas/hprov/english/hp-adm_eng.html, 2007-5-30) Normal score is the most important score to decide if the candidate could be accepted by the university he or she applies. Range of NORMAL is (0, 2.0).

2.2. Methods

2.2.1. Descriptive statistics

Descriptive statistics are used to describe the basic features of the data. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data. Descriptive statistics are typically distinguished from inferential statistics. With descriptive statistics we are simply describing what is or what the data shows. Descriptive Statistics are used to present quantitative descriptions in a manageable form. In a research study we may have lots of measures. Or we may measure a large number of people on any measure. Descriptive statistics help us to simplify large amounts of data in a sensible way. Each descriptive statistic reduces lots of data into a simpler summary.

In this paper, descriptive statistics are used to show the distribution of score, such as **box plot, bar plot,** and **mean**. A **box plot** is a convenient way of graphically depicting the five-number summary, which consists of the smallest observation, lower quartile (Q1), median, upper quartile (Q3), and largest observation; in addition, the box plot indicates which observations, if any, are considered unusual, or outliers. Box plots are able to visually show different types of populations, without any assumptions of the statistical distribution. The spacing between the different parts of the box helps indicate variance, skew and identify outliers. A **bar chart**, also known as a bar graph, is a chart with rectangular bars of lengths usually proportional to the magnitudes or frequencies of what they represent. Bar charts are used for comparing two or more values. A **mean** or average is probably the most commonly used method of describing central tendency. At the same time, **ratio scale** is also used, which is typically unit less, as they relate quantities of the same dimension. A rate is a special kind of ratio in which the two quantities being compared are of different units.

2.2.2. T-test

A **t-test** is any statistical hypothesis test in which the test statistic has a Student's t distribution if the null hypothesis is true. Test of the null hypothesis that the means of two

normally distributed populations are equal, H_o: $\mu_1 - \mu_2 = 0$, H_1: $\mu_1 - \mu_2 \neq 0$. Given two data sets, each characterized by its mean, standard deviation and number of data points; we can use some kind of t test to determine whether the means are distinct, provided that the underlying distributions can be assumed to be normal.

The t statistic to test whether the means are different can be calculated as follows:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{\bar{X}_1 - \bar{X}_2}} \text{ where } s_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \quad (1)$$

Where s^2 is the unbiased estimator of the variance, n = number of participants, 1 = group one, 2 = group two. $n - 1$ is the number of degrees of freedom for either group, and the total sample size minus 2 is the total number of degrees of freedom.

The assumption of t-test is,

A. Normal distribution of data, which could be tested by using the Shapiro-Wilk test.

B. Equality of variances, tested by using the F test.

C. Samples may be independent or dependent, depending on the hypothesis and the type of samples: Independent samples are usually two randomly selected groups; dependent samples are either two groups matched on some variable or are the same people being tested twice.

In this paper, to use t-test, I first use the Shapiro-Wilk test to test if the data is following the normal distribution.

Shapiro-Wilk test tests the null hypothesis that a sample $x_1, \ldots, x_n$ came from a normally distributed population. "The test statistic is,

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad (2)$$

Where $x_{(i)}$ is the *i*th order statistic, and a_i are given by,

$$(a_1,...,a_n) = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{\frac{1}{2}}} \qquad (3)$$

Where $m = (m_1,...,m_n)^T$, and m_1, ..., m_n are the expected values of the order statistics of independent and identically-distributed random variables sampled from the standard normal distribution, and V is the covariance matrix of those order statistics."

(http://en.wikipedia.org/wiki/Shapiro-Wilk_test, 2007-5-30)

2.2.3. Interval estimation

Interval estimation is the use of sample data to calculate an interval of possible (or probable) values of an unknown population parameter. The most prevalent forms of interval estimation are confidence intervals.

"Confidence interval (CI) for a population parameter is an interval with an associated probability p that is generated from a random sample of an underlying population such that if the sampling was repeated numerous times and the confidence interval recalculated from each sample according to the same method, a proportion p of the confidence intervals would contain the population parameter in question. If U and V are statistics (i.e., observable random variables) whose probability distribution depends on some unobservable parameter θ, and

$$\Pr(U < \theta < V \,|\, \theta) = x \text{, (where x is a number between 0 and 1)} \qquad (4)$$

Then the random interval (U, V) is a "(100·x) % confidence interval for θ". The number x (or 100·x %) is called the confidence level or confidence coefficient."

(http://en.wikipedia.org/wiki/Confidence_interval, 2007-5-30)

In this paper, confidence interval is used to find if there are some specific questions that could cause gender bias.

3. Results

In this part, there are three parts, descriptive analysis, advanced data analysis, and analysis of the three questions mentioned in the introduction, is there gender difference? Is there some specific tests? And is there any specific questions?

3.1. Descriptive analysis

First, to have a look at all data, a box plot (Figure 1) is used to show the distribution of scores of five subtests, total score, and normal score.

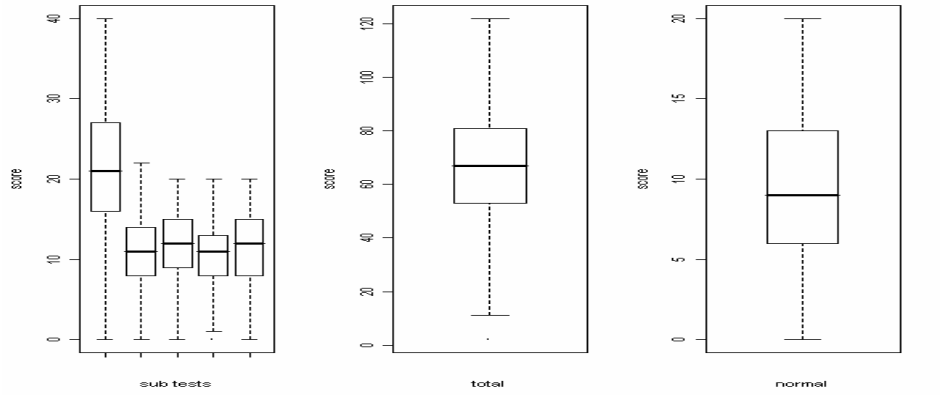

Figure 1. Box plot of sub-tests, total, and normal score for all ten test-occasions. The sub-tests are (from the left) ORD, NOG, LAS, DTK, and ELF.

Then, split the scores of five subtests, total score, and normal score by gender and get the box plot for male and female separately (Figure 2) .

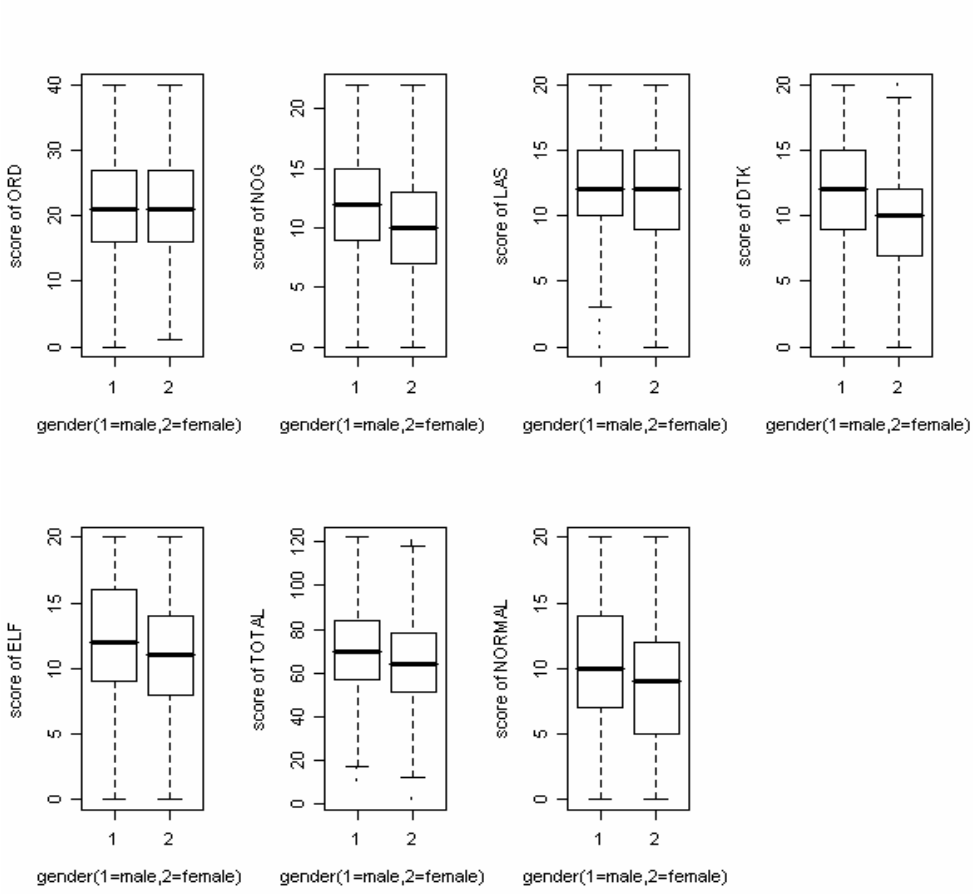

Figure 2. Box plot of subtests, total, and normal score for all data by gender.

It is obviously that the scores of male are higher than scores of female in most of the test results, especially for NOG, DTK, ELF, TOTAL, and NORMAL. At the same time, we can see the means of the scores in table 2.

Table 2. Means of subtests, total, and normal score for data of all ten test-occasion (Spring 2002 to autumn 2006) by gender and the difference of male and female.

	Male	**Female**	**Difference**
ORD	21.7	21.5	0.3
NOG	12.2	10.3	1.9
LAS	12.4	12.1	0.3
DTK	11.8	9.9	1.9
ELF	12.3	11.1	1.2
TOTAL	70.3	64.8	5.5
NORMAL	10.1	8.7	1.4

As table 2 shows, scores of male are always higher. To find that if this is coursed by random or there is really gender bias in the SweSAT, I have to do some advanced data analysis.

3.2. Advanced data analysis

As table 2 shows, there are differences between male and female in all data. So to know if there is gender difference, the data file have to be split into 10 sub data file by time, from spring 2002 to autumn 2006, and then, analyze every sub data files. Here, take the result of spring 2006 as an instance.

3.2.1. Mean by gender

For a data set, the mean is the sum of the observations divided by the number of observations. The mean describes the central location of the data, is the average of all data.

Table 3 is the table of means for all data and by gender of data file of spring 2006. In table 3, the central location of scores for the tests can be presented.

Table 3. Table of means of data file of spring 2006, including the total mean (Total), mean of male (Male), mean of female (Female), and the difference of male and female (male-female) (Difference).

Name	Total	Male	Female	Difference
ORD	21.6	22.3	21.0	1.3
NOG	11.4	12.5	10.4	2.1
LAS	13.0	13.1	12.9	0.2
DTK	11.2	12.2	10.4	1.9
ELF	11.2	12.0	10.5	1.5
TOTAL	68.3	72.1	65.1	6.9
NORMAL	9.3	10.3	8.5	1.7

3.2.2. Normalized mean (Correct ratio)

Since the maximum points of each sub-test are not the same, it can not be compared

directly. To make the differences comparable, the mean of the number of correct answers have to be normalized, using the correct ratio instead of using the raw data.

The formula of normalized mean is the proportion of correct answer,

$$Normalized\ mean = \text{Prob}(correct\ answer) = \frac{mean}{No.\ of\ questions} \qquad (5)$$

Take ORD as an example; from table 3, we can see the mean of ORD is 21.56, and there are 40 questions in this subtest, so the normalized mean of this is,

$$Total(\text{ORD}) = \text{Prob}(correct\ answer \mid sub\ test = \text{ORD}) = \frac{21.56}{40} = .539 \qquad (6)$$

Then the table of normalized mean (table 4) can be computed and let the data comparable.

In table 4, it is shown that there are two tests with largest normalized difference, NOG and DTK. So in next step, we will pay attention to these two parts.

Table 4. Table of normalized means of data file of spring 2006, including the normalized total mean (Total), normalized mean of male (Male), normalized mean of female (Female), and the difference of normalized means of male and female (male-female) (Difference) for five sub-tests and total scores.

Name	Total	Male	Female	Difference
ORD	0.539	0.557	0.524	0.033
NOG	0.516	0.567	0.474	0.093
LAS	0.648	0.654	0.643	0.011
DTK	0.561	0.612	0.519	0.093
ELF	0.560	0.600	0.526	0.074
TOTAL	0.560	0.591	0.534	0.057

3.2.3. Analysis of each question for the subtest with largest difference of mean by gender

Now, pay attention to NOG and DTK to identify if there are specific questions causing the gender difference. In figure 3 the normalized differences are presented for each question in the two sub-tests.

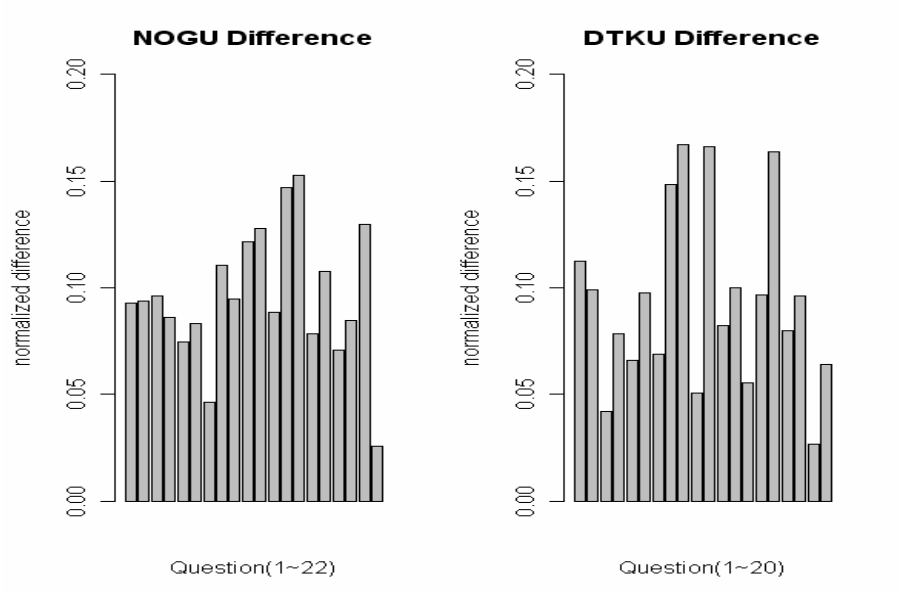

Figure 3. Bar plots of the normalized difference for each question in the two sub-tests (NOG and DTK) of spring 2006.

In this figure, all questions seem to indicate gender bias.

3.3. Three questions

3.3.1. Is there gender difference?

To answer this question, I pick up 10 **normal differences** (Normal difference=Mean of normal score for male-Mean of normal score for female) from tables of mean by year. We can see the result in table 5.

Table 5. Normal difference by year.

Test-occasion	Year	Normal difference
1	Spring 2002	1.03
2	Autumn 2002	1.27
3	Spring 2003	1.42
4	Autumn 2003	1.49
5	Spring 2004	1.08
6	Autumn 2004	1.46
7	Spring 2005	1.41
8	Autumn 2005	1.37
9	Spring 2006	1.72
10	Autumn 2006	1.35

3.3.1.1. Shapiro-Wilk Normality Test

To test if the "normal difference" is following normal distribution, I use Shapiro-Wilk Normality Test. The null hypothesis is: the sample came from a normally distributed population. The result shows that W=0.9445 and p-value of Shapiro.test is equal to 0.6046, so the null hypothesis can not be rejected, which means that the distribution of data is not significantly different from the normal distribution and t-test can be used in the next step.

3.3.1.2. T-test of normal difference

The other assumption of using t-test is that the data are independent. In fact, there are some candidates take the *(n+1)*th test who have taken the *n*th. So I can not say that the 10 data are independent. But the proportion of the candidates of retest is about 20%, so here I can use approximately t-test.

The t-test statistic is t=21.5574 with df = 9, and the p-value is equal to 4.682e-09, less than 0.05, which will not allow us to accept the null hypothesis，the mean is equal to 0, $\mu=0$.

3.3.1.3. Result

Since the null hypothesis is: the mean is equal to 0, μ=0, we can say that there is gender difference in the SweSAT.

3.3.2 Is there some specific tests?

To answer this question, the differences of normalized means (correct ratio) of five subtests of ten tests were picking up. The p-value of Shapiro-Wilk Normality Test are all larger than 0.05, which shows that all the data are following normal distribution and t-test can be used to test the mean.

3.3.2.1. T-test of normalized difference.

In table 6, the differences of normalized mean of five sub-tests are present. The correct ratio of male is always larger than the correct ratio of female.

Table 6. Difference of normalized mean of five sub-tests of ten test-occasions.

Test-Occasion	ORD_N_D	NOG_N_D	LAS_N_D	DTK_N_D	ELF_N_D
1	-0.015	0.097	0.004	0.104	0.07
2	0.007	0.073	0.014	0.084	0.059
3	0.016	0.085	0.003	0.097	0.057
4	0.003	0.096	0.023	0.103	0.052
5	-0.003	0.075	0.014	0.08	0.044
6	0.008	0.076	0.018	0.102	0.063
7	0.003	0.087	0.024	0.094	0.063
8	0.003	0.084	0.025	0.089	0.055
9	0.033	0.093	0.011	0.093	0.074
10	0.007	0.087	0.002	0.097	0.059
Mean	0.0062	0.0853	0.0138	0.0943	0.0596

Now, use t-test on the data of table 6. The null hypothesis is: the mean is equal to 0, μ=0. The result of t-test is in table 7.

Table 7. Result of t-test.

	t	df	p-value
ORD	1.58	9	0.1483
NOG	31.45	9	1.63E-10
LAS	4.98	9	0.000761
DTK	37.02	9	3.80E-11
ELF	21.87	9	4.12E-09

In table 7, the p-values are all less than 0.05 except the p-value of ORD.

3.3.2.2. Result

After the analysis, it is obvious that there are gender differences except for "ORD". The gender biases are shown in figure 4

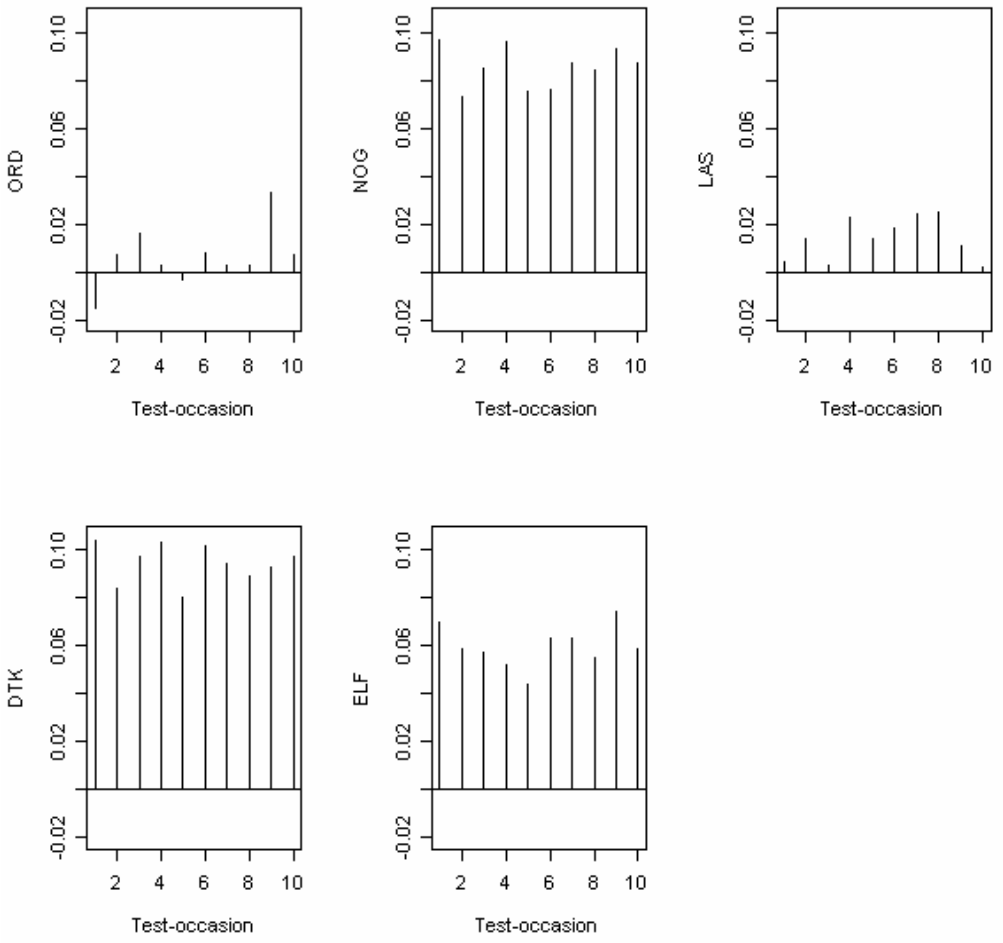

Figure 4. Plot of difference of normalized mean of five sub-tests of ten test-occasions.

3.3.3. Is there any specific questions?

In the normalized mean table (table 6), the largest difference of means between Male and Female is the difference of DTK and NOG. So start from the subtest with large difference of mean to the subtest with small difference of mean to identify if there are specific questions causing gender difference. In this part, data of autumn 2006, the latest one will be focus on.

3.3.3.1. Analysis of five subtests.

Even there are five subtests; they can be divided into two parts by the result of t-test to every question of each subtest. The first part including three subtests, DTK, NOG, and ELF, in which almost all questions have gender bias, result of male is better than result of female. The second part including the other two subtests, ORD and LAS, in which the result of some questions shows that male is better than female as well as the result of some other questions shows that female is better than male.

3.3.3.1.1. The first part.

3.3.3.1.1.1. Test of DTK, the subtest about interpretation of diagrams, tables, and maps.

In this part, the main method is confidence interval. First, get the upper bound, lower bound of confidence interval from the difference in the proportion of correct answers, and the estimate (table 8).

Table 8. Upper bound, lower bound of confidence interval (the difference in the proportion of correct answers), and the estimate of every question of autumn 2006.

Question	Upper Bound	Estimate	Lower Bound
1	0.067	0.078	0.089
2	0.027	0.037	0.046
3	0.130	0.141	0.152
4	0.055	0.065	0.076
5	0.012	0.023	0.034
6	0.070	0.081	0.092
7	0.026	0.037	0.048
8	0.131	0.142	0.153
9	0.152	0.163	0.174
10	0.131	0.143	0.154
11	0.158	0.169	0.180
12	0.062	0.073	0.084
13	0.097	0.108	0.119
14	0.090	0.102	0.113
15	0.106	0.117	0.128
16	0.090	0.101	0.112
17	0.158	0.169	0.180
18	0.040	0.050	0.061
19	0.070	0.080	0.091
20	0.046	0.057	0.068

To know the distribution straightforward, figure 5 is used to show it.

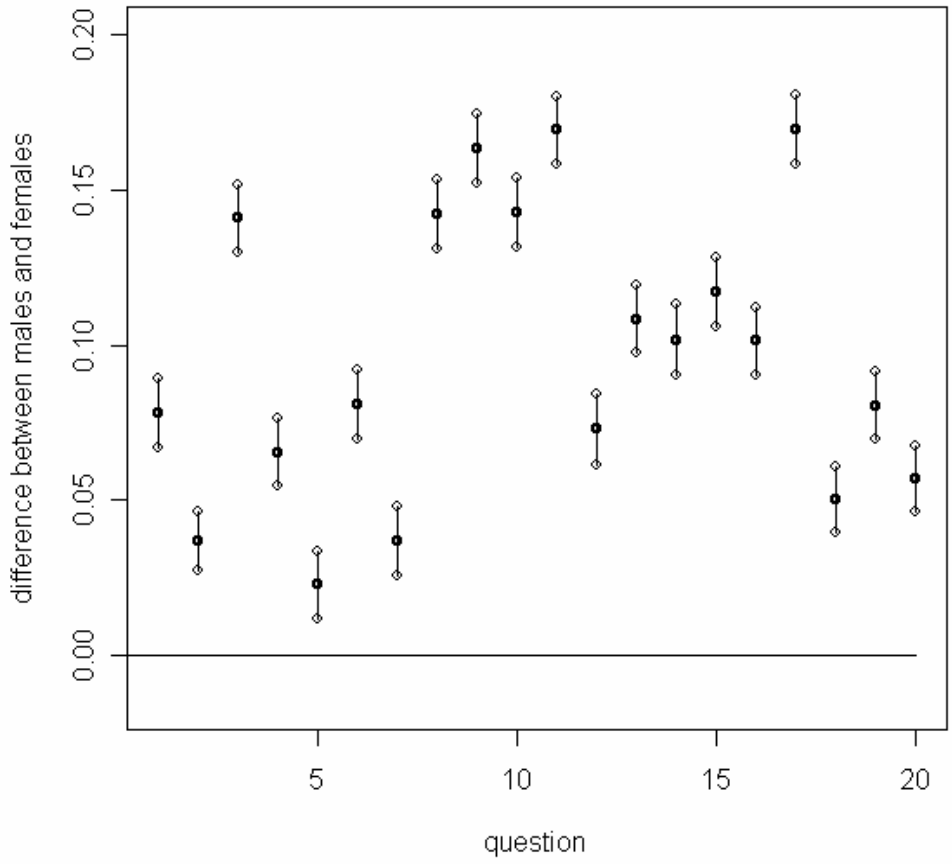

Figure 5. Plot of confidence interval of DTK of autumn 2006.

In figure 5, confidence intervals of all the 20 questions are significant, which means that the mean of male and female are not the same for every question. As figure 5 shows, confidence interval of every question are all above zero line, which means that the mean of male of the 20 questions are all larger than the mean of female. The result can be also getting by p-value of t-test.

3.3.3.1.1.2. Test of NOG, the subtest about data sufficiency.

Use the same method of the analysis of DTK, figure 6 can be computed, the plot of confidence interval of sub-test NOG.

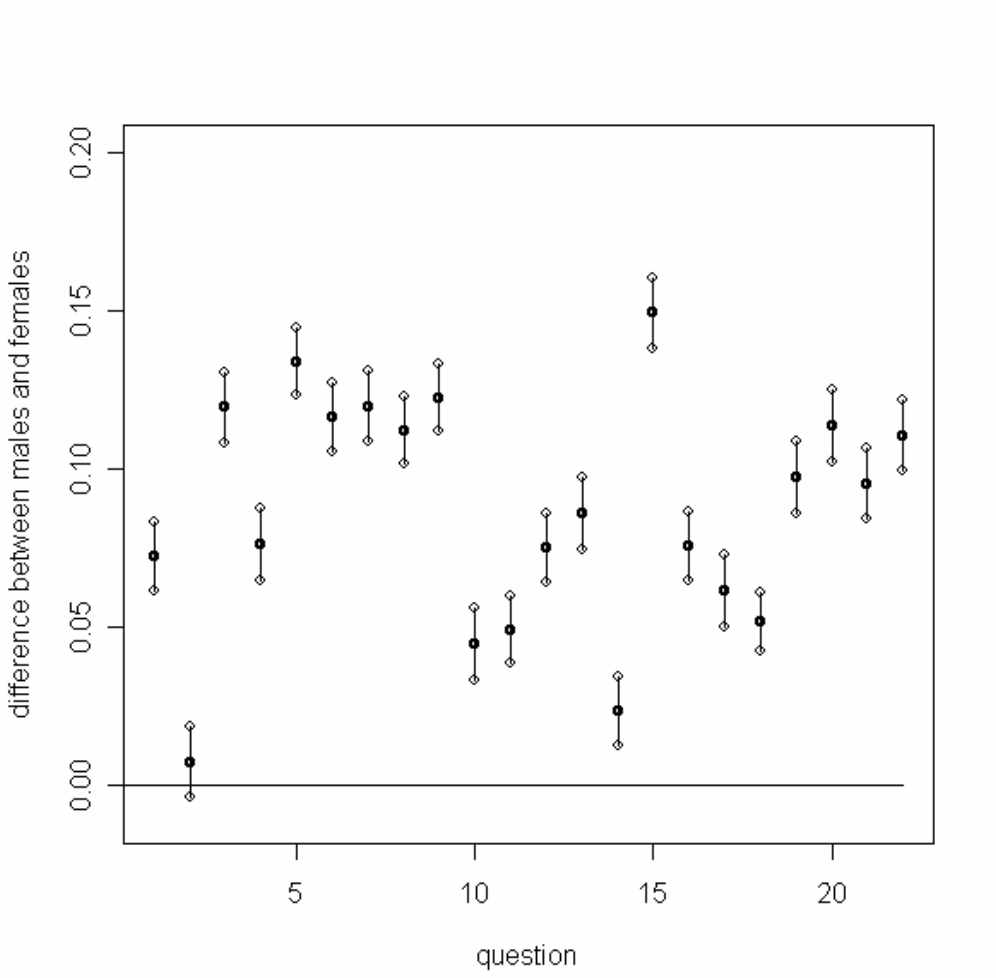

Figure 6. Plot of confidence interval of NOG of autumn 2006.

In figure 6, confidence interval of all the 22 questions except the second question are significant, which means that the mean of male and female are not the same for every question except the second question. As figure 6 shows, confidence interval of every question except for the second question are all above zero line, which means that most of the mean of male of the 22 questions are larger than the mean of female. For the second question, since it is just on the zero line, the null hypothesis can not be rejected; the mean of male is the same as the mean of female.

3.3.3.1.1.3. Test of ELF, the subtest about English reading comprehension.

Also use the same method of the analysis of DTK, figure 7 can be computed, the plot of

confidence interval of sub-test ELF.

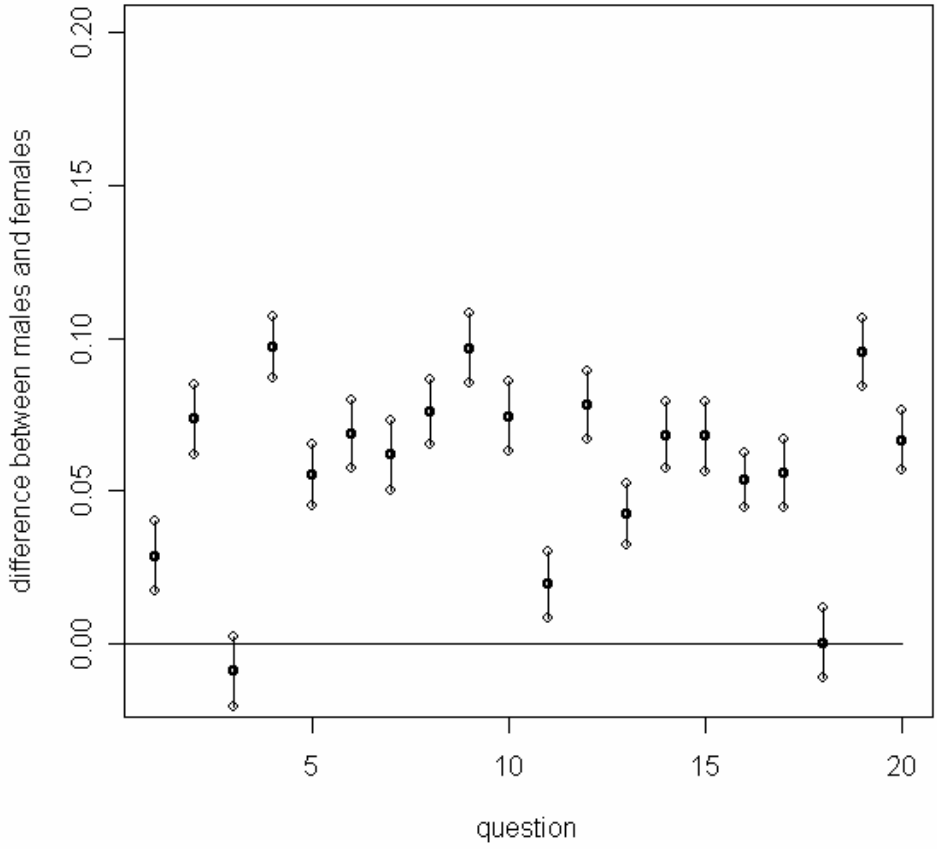

Figure 7. Plot of confidence interval of ELF of autumn 2006.

In figure 7, confidence interval of all the 20 questions except the third and the eighteenth questions are significant, which means the mean of male and female are not the same for every question except for the third and the eighteenth questions. As figure 7 shows, confidence interval of every question except for the third and the eighteenth questions are all above zero line, which means that most of the mean of male of the 20 questions are larger than the mean of female. For the third and the eighteenth questions, since they are just on the zero line, the null hypothesis can not be rejected; the mean of male is the same as the mean of female.

3.3.3.1.2. The second part.

For this part, some advanced methods are added to analyze the result of t-test because of the different distribution of confidence interval.

3.3.3.1.2.1. Test of ORD, the subtest about vocabulary.

First, do the same thing as the first part, get figure 8.

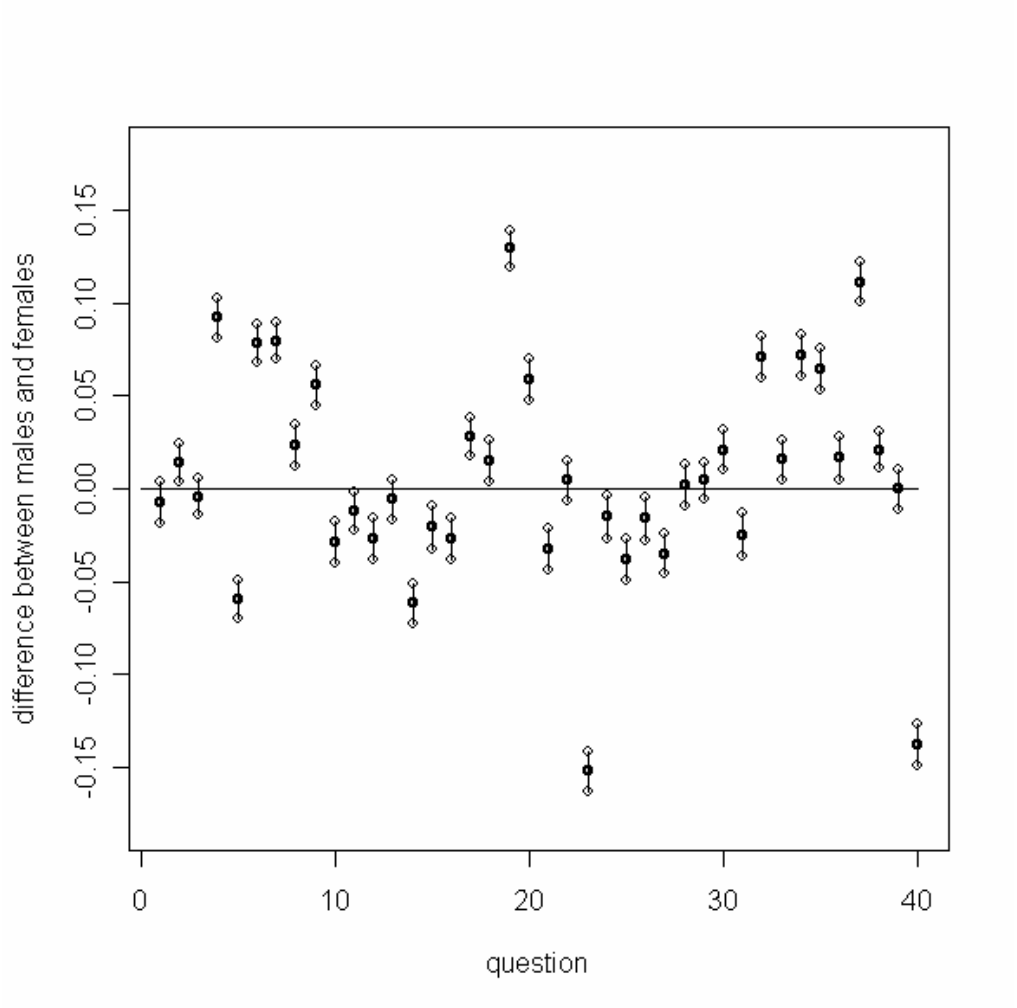

Figure 8. Plot of confidence interval of ORD of autumn 2006.

From figure 8, we can see that 7 out of 40 questions are not significant, which means that the null hypothesis is accepted, the mean of male and female are almost the same for these 7 questions. For the other 33 questions, there are 18 confidence intervals above zero

line, whereas 15 confidence intervals below zero line. In the other word, we can see the relationship clearly in figure 9.

$$\left\langle \begin{array}{l} Significant\_questions = 33 \left\langle \begin{array}{l} questions\_above\_zero = 18 \\ questions\_below\_zero = 15 \end{array} \right. \Rightarrow \left\langle \begin{array}{l} mean(male) > mean(female) \\ mean(male) < mean(female) \\ mean(male) = mean(female) \end{array} \right. \\ Non-significant\_questions = 7 \end{array} \right.$$

Figure 9. Relationship of 40 questions of ORD of autumn 2006.

In figure 9, the amount of questions (mean (male)>mean (female)) is close to the amount of questions (mean (male) <mean (female)). So for this subtest, there is not significant gender bias.

3.3.3.1.2.2. Test of LAS, the subtest about reading comprehension.

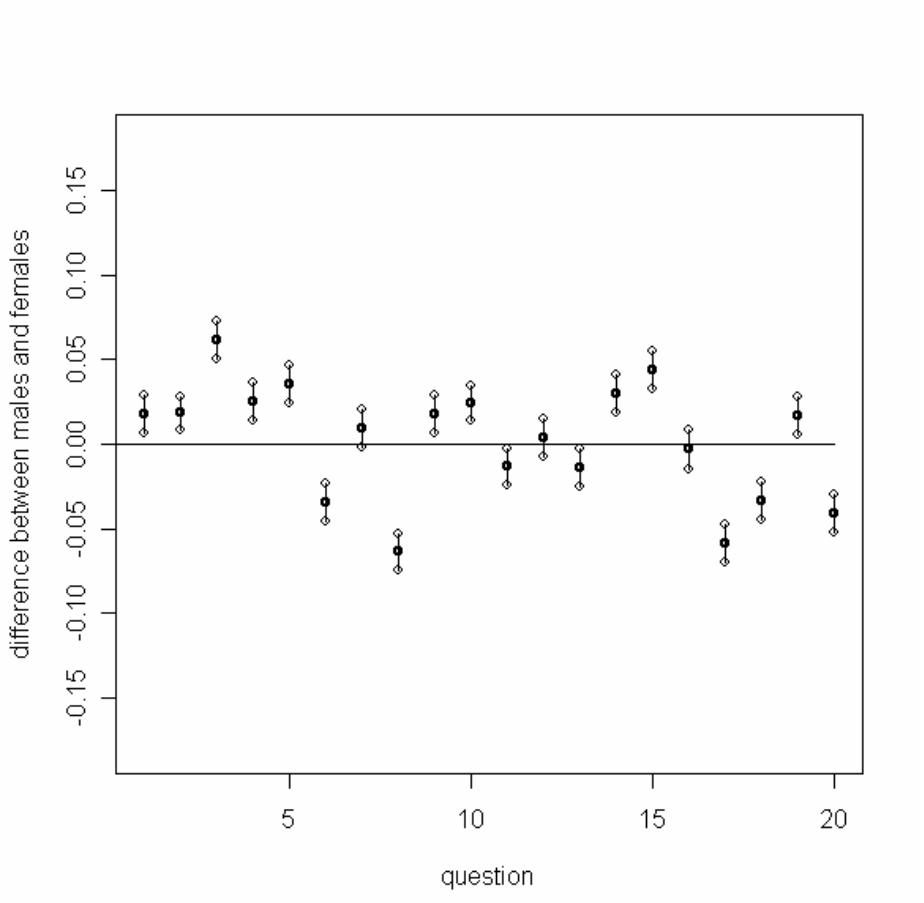

Figure 10. Plot of confidence interval of LAS of autumn 2006.

From figure 10, we can see that 3 out of 20 questions are not significant, which means that the null hypothesis is accepted, the mean of male and female are almost the same for these 3 questions. For the other 17 questions, there are 10 confidence intervals above zero line, whereas 7 confidence intervals below zero line. In the other word, we can see the relationship clearly in figure 11.

$$\left\langle \begin{array}{c} Significant\_questions = 17 \left\langle \begin{array}{l} questions\_above\_zero = 10 \\ questions\_below\_zero = 7 \end{array} \right. \\ Non-significant\_questions = 3 \end{array} \right. \Rightarrow \left\langle \begin{array}{l} mean(male) > mean(female) \\ mean(male) < mean(female) \\ mean(male) = mean(female) \end{array} \right.$$

Figure 11. Relationship of 20 questions of LAS of autumn 2006.

In figure 11, the amount of questions (mean (male)>mean (female)) is close to the amount of questions (mean (male) <mean (female)). So for this subtest, there is not significant gender bias.

This result contradicts the previous conclusions e.g. table 7, where shows that the LAS test has gender bias. The interpretation of the LAS results are a bit tricky. On average there seems to be a small gender bias even though some questions do favor women. Hence, by another selection of questions it should be possible to cause gender bias in favor for women.

3.3.3.2. Result

After the analysis by method of confidence interval, we can get the result. For the first part, the subtests of DTK, NOG, and ELF, almost all questions are specific questions causing gender difference, and for the second part, the subtests of ORD and LAS, the gender difference is not significant.

4. Conclusions

After the analysis, the three questions mentioned as the aim of this paper can be answered.

First, is there a gender difference? Yes, there is gender difference in the Swedish Scholastic Assessment Test.

Second, are there some specific tests? Yes, 4 out of 5 subtests have significant gender difference, except for the sub-test of ORD.

Third, are there some specific questions? Yes, all of questions of the subtests of DTK are significant, and most of questions of the subtests of NOG and ELF are significant to lead to gender bias. For sub-tests of ORD, there is not significant gender bias, and for the sub-tests of LAS, there is gender bias when using the method of t-test but there is not significant gender bias when using the method of confidence interval.

By analyzing the ten test-occasions data files, the results shows that there is gender bias in the SweSAT. Is it OK to have gender bias? Or is there discriminating against women? It is discussed that a possible explanation of this situation is that smart women get into university with their school grades and they therefore don't have to write the SweSAT test while smart boys get worse grades in school and hence needs to rely on this test. This could be one of the reasons of gender bias in the SweSAT. But at the same time, the existing of gender bias also shows that there is something to be improved of the SweSAT to decrease the differences between male and female.

References

1. Brain S. Everitt, Hothorn T. (2006), A Handbook of Statistical Analyses Using R, Chapman &Hall/CRC, Taylor & Francis Group

2. Carter M., Williamson D. (1996), Quantitative Modelling for Management & Business, Pitman Publishing

3. Casella G., Roger L. Berger (2002), *Statistical Inference, Second Editio*n, Duxbury, Thomson Learning

4. Dalgaard P. (2002), *Introductory Statistics with* R, Springer-Verlag New York, Inc.

5. David R. Anderson, Dennis J. Sweeney, Thomas A. Williams, *Statistics for Business and Economic*s, Fifth Edition, West Publishing Company

6. Department of Educational Measurement, Faculty of Social Sciences, Umeå University. The Swedish Scholastic Assessment Test, SweSAT. Available on line at http://www.umu.se/edmeas/hprov/index_eng.html, (2007-5-13)

7. Department of Educational Measurement, Faculty of Social Sciences, Umeå University. Available on line at http://www.umu.se/edmeas/hprov/english/hp-adm_eng.html, (2007-5-30)

8. Wikipedia, the free encyclopedia. Confidence interval. Available on line at http://en.wikipedia.org/wiki/Confidence_interval, (2007-5-30)

9. Wikipedia, the free encyclopedia. Shapiro-Wilk test. Available on line at http://en.wikipedia.org/wiki/Shapiro-Wilk_test, (2007-5-25)

Appendix

Appendix 1.

Table 1. Names of variables

Old names	New names
ID	ID
PROVAR	YEAR
PROVTILLFALLE	SEASON
FODELSEAR	BIRTH
PAGAENDE_SVENSK_UTBILDNING	ONGOING
AVSLUTAD_SVENSK_UTBILDNING	FINISH
ALDER	AGE
KON	**GENDER**
MAX_SVENSK_UTBILDNING	MSU
PAG_SVENSK_UTBILDNING	PSU
AVS_SVENSK_UTBILDNING	ASU
UTLANDSK_UTBILDNING	UU
ORD	**ORD**
NOG	**NOG**
LAS	**LAS**
DTK	**DTK**
ELF	**ELF**
TOTALPOANG	**TOTAL**
NORMERAD_POANG	**NORMAL**
ORD_Uppg_1-40	**ORDU**
NOG_Uppg_1-22	**NOGU**
LAS_Uppg_1-20	**LASU**
DTK_Uppg_1-20	**DTKU**
ELF_Uppg_1-20	**ELFU**