Department of Economics and Society, Dalarna University

STATISTICS
Master's Thesis D 2007

Effect of repeated testing in SweSAT for entrance to university

Examiner: **Prof. Kenneth Carling**

Supervisor: **Prof. Johan Bring**

Authors: **He Zhang** and **Sanja Kukolj**

ABSTRACT

The main purpose of this paper is to analyze the effects of repeated Swedish Scholastic Aptitude Test (SweSAT) taking. SweSAT is a selection test used for entrance to university in Sweden with aim of measuring the test takers' expected ability for higher education as fairly as possible. SweSAT is offered twice per year. This research covers period of five years from 2002 to 2006, in both April and October terms. The test consists of 122 items. Total score of 122 is transformed to a normalized score and used in the selection procedure. Results of ten tests are sufficient to give a good description of effects of test repeating in terms of normalized score and probability of improvement.

Additionally the possibility of taking test several times according to the rule of SweSAT will give test-takers a chance for improving their results. The general conclusion is that most of the test takers take the test twice and that the biggest improvement occurs between the first and second test. The more tests they take the less chance of improvement they have. However there is learning effect existing during test repeating so that the probability of improvement is larger in real data than that in empirical data. Improvement in each subtest varies from test to test, but the biggest improvement can be made in the WORD test. Men have larger probability of improvement in test results than women. The differences between each age group have no regularity. However, there is still a trend indicating that younger people have a bit larger probability of improving.

Key words: SweSAT, Normalized Points, Probability of Improvement

TABLE OF CONTENTS

1. INTRODUCTION

Swedish Scholastic Aptitude Test (SweSAT) was introduced, by Agency for Higher Education, in spring 1977 as one way of selection procedure to higher education in Sweden. "The SweSAT is supposed to measure acquired (developed) abilities and makes use of the kind of verbal and mathematical skills that develop over the years, both in and outside school. The content of the test does not reflect any specific curriculum, although it is designed to be consistent with school-based learning." [1] "Selection can also be made by grade point average from upper secondary school (GPA) or by scores from various types of course-specific selection instruments for programs like architecture, journalism and medicine." [2] "SweSAT is regarded as a second chance for applicants who have grades, since they are automatically judged on the most favorable condition, and do not have to decide by the applicants whether to use the grades or test results." [3]

Since 1996 SweSAT has developed as a test with 122 multiple choice items divided into five subtests. Forty items in subtest WORD are testing understanding of vocabulary and concepts (both Swedish and loan-word). Twenty-two items in subtest DS (Data Sufficiency) are measuring test takers' logically reasoning ability. Subtest READ (Swedish Reading Comprehension) includes twenty items corresponding to five passages of one-page length, which tests the comprehension of the text as a whole or details stated. Solving the subtest DTM in ten figures and twenty questions, test takers should show ability to read and understand figures, tables and maps. Subtest ERC (English Reading Comprehension) consists of eight to ten texts varies in length and in item format. Applicants are supposed to answer twenty questions within 35 minutes. The total testing time is four hours and ten minutes. (For WORD test 15 minutes and for tests DS, READ and DTM 50 minutes. In addition, there is a set of pre-test items which lasts for 50 minutes and it is not included in the test-taker's result.)

Total scoring is done by summing up all the points from these five subtests (122 points in total with each item having one point). However, the test on each occasion is not of the same difficulty. To make test results comparable, normalized point is calculated by standardizing total point into the interval between 0.0 and 2.0 (0-20 are codes used to denote the real interval in the analysis part of this essay). "Test taker's best result of normalized score is competed for university admission and the test result is valid for five years. Beside this, other requirements should be met: it should be in line with the aims of higher education, it should not have negative effects on courses preceding it, it should be possible to mark quickly, inexpensively and

objectively, and it should not be possible to improve test results by mechanical exercises or by learning special principles for solving problems. The test should be experienced as meaningful by the test takers and no group should be favored or discriminated against because of gender or social class."[3]

Knowing these entire facts questions rise: whether repeating tests makes sense and whether there is any learning effect during test repeating or not. What is the probability of improvement through test repeating? Is there any difference between gender and age groups in terms of normalized point or chance of improvement?

The paper is divided into four parts: Section 2 describes how the data are grouped in terms of repetition. The grouped data are the base of the following analysis. Section 3 mainly analyze the effects of test-repeating, that includes five sub-analyses: i) Comparison of number of test takers in each subgroups and their test results; ii)Comparison of the probability of improvement in total points between results derived in real data and R simulation; iii)Effects of test-repeating on probability of improvement in five subtests, total and normalized points for each repeating group; iv)Effects of repeating in change of total and normal points; v) Group (gender, age) comparison of normalized points in different repeating times. The discussion is done in the last section.

2. BRIEF DESCRIPTION OF DATA ANALYSIS AND METHOD USED

Data given for period 2002-2006 are first organized and imported into the R program where the analyses are done. "The reason for selecting five-year period including 10 tests results is that it allows a description of effects of test repeating at the same time as the possibilities of score changes as a function of true changes of test taker's ability are controlled for, or at least minimized." [4]

Every year there were on average 73680 people who took SweSAT, with 57% of them took the test in spring, 43% in autumn. The average proportion of repeaters at each test occasion was around 37%, which is obtained by restricting 2002 and 2003 as control years. (See *Appendix B*, Table 1) The proportion of repeaters in control years are not calculated because whether the test takers took the test before 2002 is unavailable. Each proportion of repeaters in the tests offered in 2004 to 2006 is derived by checking if the test takers had already taken test before (from 02A to

the time when the test is checked). (2002-2006 are denoted as 02-06, spring and autumn are denoted as A and B respectively).

The data include general information of the test-takers (ID (identification number), gender, age, education level and all the tests results, such as if the question is right or wrong, points obtained in each subtest, total and normalized points). For the analysis of the effect of repeating SweSAT, test-taker's ID and each test results including scores of WORD, DS, READ, DTM and ERC, total point and normalized point are picked up firstly through R commands. ID number is used to order the data and to see who repeated the test and how many times during the period of 2002-2006. Grouping test takers by the fact how many times they took SweSAT, we end up with 10 test-taking groups which are independent from each other. "A test taker belongs to test taking group 4", means that he/she took the test only 4 times during the period 2002 to 2006 and this test taker can not be in any other group.

Similarly, test taking groups 2 to10 can be also named as test repeating groups according to the times of repeat. The final result of a test taker in each repeating group is compared to the maximum of his/her previous one which was used in the last selection competition to see if the best result can be improved or not. Repeating groups focus on the analysis of the final result of repeating the test only i times, excluding the effect of the 1^{st} to $i-1^{th}$ repeating ($i=1,...,9$). However, the aim of this essay is to explore the effect of repeating. Therefore, the analysis of the effect of all i times of repeating is done. By comparison of the results, it is concluded that the i^{th} test repeating group generally represents the effect of test repeating i times during the years 2002-2006.

Therefore the following analysis of effects of repeating is based on each repeating group, which includes division of subgroups (by gender, age and test result), comparison of repeating effects in subgroups and comparison of that between real and empirical data. Empirical result is simulated in R program.

3. ANALYSES OF THE EFFECTS OF TEST REPEATING

3.1 STATISTIC DESCRIPTION OF NUMBER OF SUBGROUPS AND THEIR TEST RESULTS

3.1.1 COMPARISON OF THE PREFERENCE OF TEST REPEATING BETWEEN SUBGROUPS

Divisions of subgroups are done in each repeating group. Subgroups include gender groups (men/women), five age groups (-20; 21-24; 25-29; 30-39; 40-) according to the given data. There is no obvious preference of test repeating between gender groups although there is a small trend (increasing proportion of men) indicating that more men than women repeating the test, which is shown in *Appendix A* Figure 1. Similarly, there is no preference of test repeating in age groups either, although the test takers who are over 40 did not repeat the test several times. The big change of proportion happens in repeating groups 8 and 9 due to the insufficiency of the data. (See Figure 2 in *Appendix A*)

Furthermore, Normal points are transformed from the total 122 points into the interval [0.0, 2.0]. Five Normalized point (NP) subgroups are made on each test repeating group, with mean value of the normalized scores of each test taker being the standard of division. Mean value of the NP is used in the analysis instead of maximum NP because mean value is more reliable in a sense that it can show a test taker's real abilities. While maximum points can happen to be just exceptional case among all the test repeating. In addition, it is interesting to compare the repeated effects among test takers with different levels of academic achievement. In this essay NP groups are those whose average normalized points belonging to the intervals [0, 4], (4, 8], (8, 12], (12, 16] and (16, 20] respectively. Those five NP groups correspond to five levels of test results.

Table 2 in *Appendix B* shows the number of five NP subgroups in each test repeating group, their total amount and corresponding proportions. It can be seen that there are relatively few people who do the test extremely bad or extremely good. The largest group represents test takers who took the test twice (67%). We also consider the preference of different NP groups on the times of repeating, which is shown in Figure 3. People who achieved low score have a decreasing proportion in different repeating groups and they would not like to repeat more than 7 times. Those who done the test on average and those above the average prefer to be persistent in repeating the test. There is a climbing trend in proportion of those test takers who achieved above the average level as times of repeating increases. Those who got higher scores have stable small proportion in each repeating group.

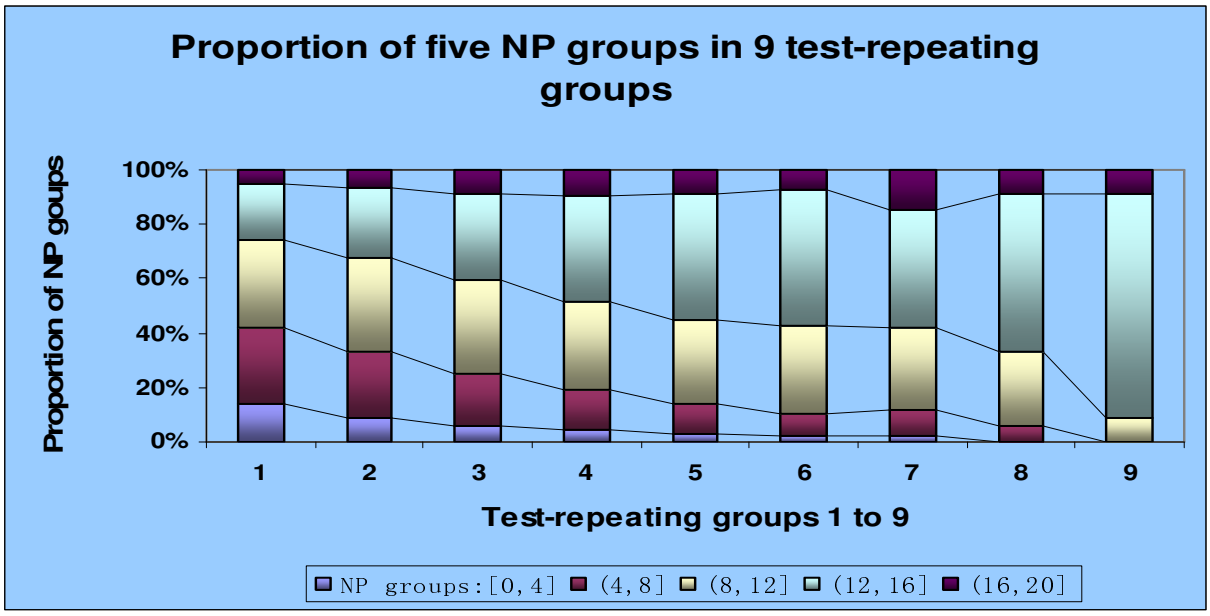

Figure 3: Proportion of 5 NP subgroups in each repeating group according to achieved mean score (normalized points) in all tests

3.1.2 COMPARISION OF TOTAL AND NORMAIZED POINT OBTAINED ON EACH TEST OCASSION

Figure 4 shows that the average normalized score for test takers who took the test only once is 0.89. If a test taker took the test several times, his/her normalized point has increased largely. The same trend can be obtained in the analysis of maximum total points which is shown in Figure 5.

In addition, comparison between gender and age groups in terms of maximum normalized points is done in Figure 6 and Figure 7(see *Appendix A*). On average women gained lower normal points than men in each test taking group; however, both of them have improved scores by repeating the test. “Females get higher grades in all school subjects except physics and physical education. This fact also contributes to the gender differences on SweSAT, since females with high grades usually do not take the test, while males with high grades take the test for competitive reasons. Since selection to higher education to a great extent is based on grades from upper secondary school, on the whole females are favored by the selection system. On average, 56 % of the new students each year are females.” [3]

The general trend for age groups comparison is that the improvement for young people was larger than that for old people as times of repeat increases. The differences of maximum normalized scores among five age groups were irregularity in each test taking group. (Figure 7 in *Appendix A*)

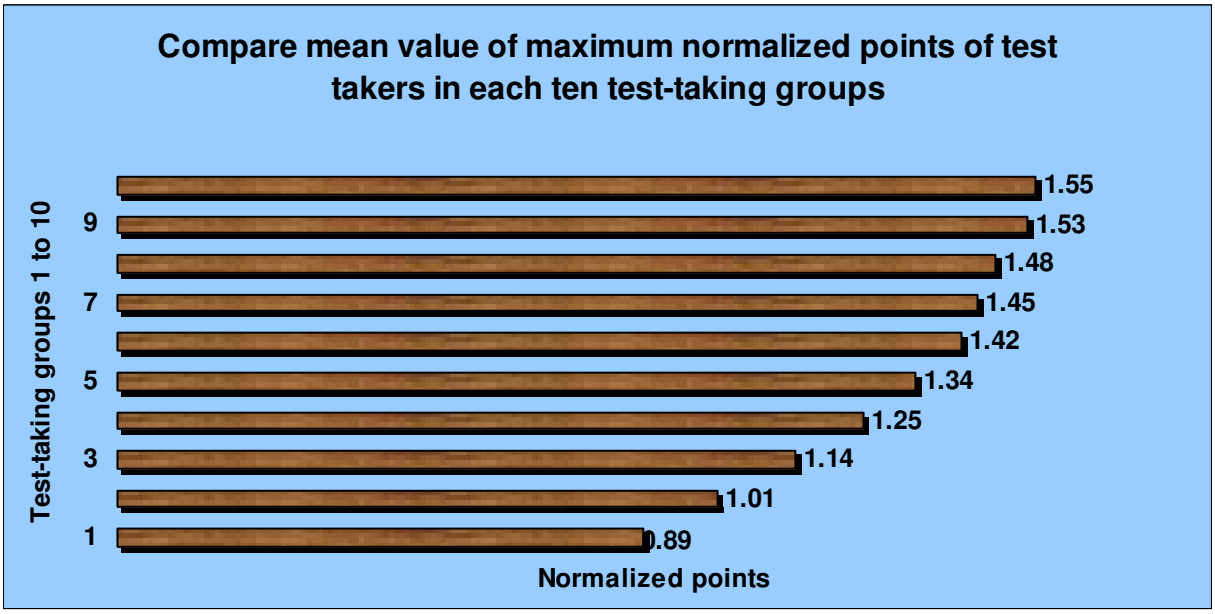

Figure 4: Compare the average value of maximum normal points in test-taking groups 1 to 10

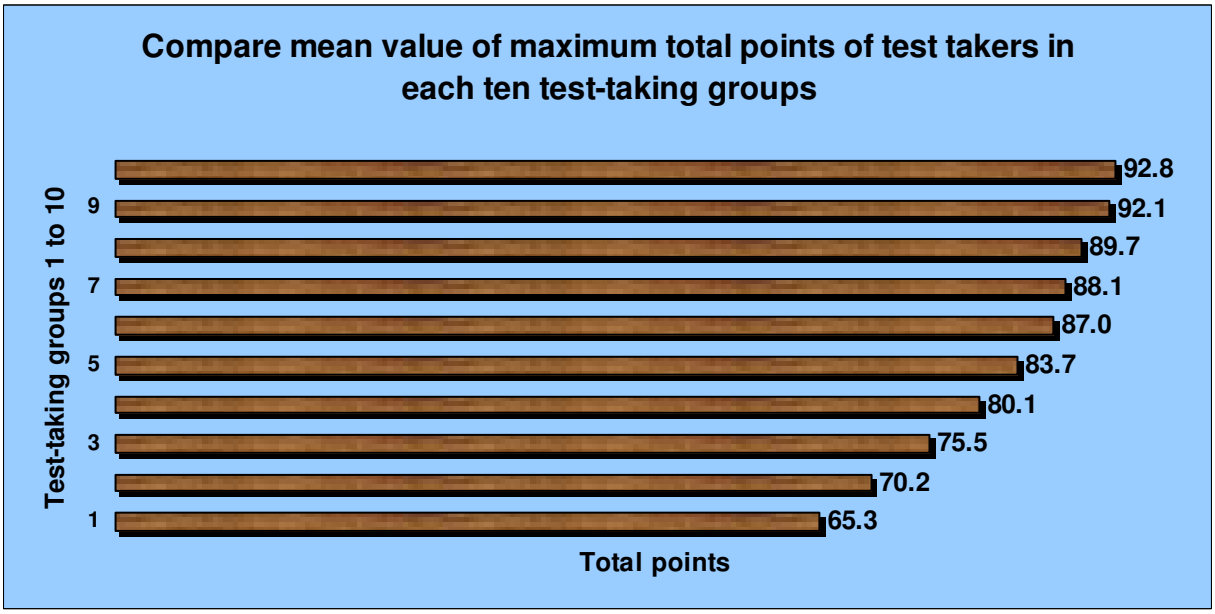

Figure 5: Compare the average value of maximum total points obtained in each test-taking group

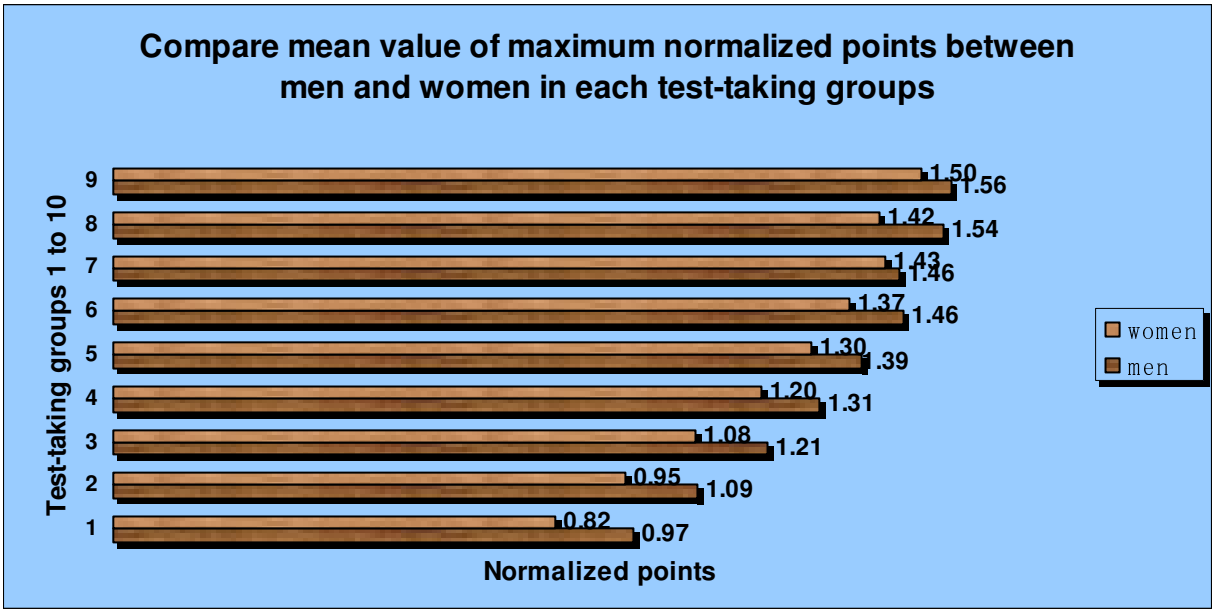

Figure 6: Compare the average value of maximum normalized points between men and women in each test-taking group

3.2 COMPARISON OF THE PROBABILITY OF IMPROVEMENT IN TOTAL POINTS BETWEEN REAL AND EMPIRICAL DATA

3.2.1 PROBABILITY OF IMPROVEMENT IN TOTAL POINT AMONG FIVE NP GROUPS

This part concentrates on comparing the probability of improvement for test takers with different levels of test result. Each probability of improvement in total point is done by comparing the final total point of a test taker with his/her best result before. First the comparison results among five NP groups from real data is analyzed and then compared with those obtained in the simulation.

Improvement should be that the last obtained result is larger than the maximum of the previous ones for each test taker. If there is an improvement, it will be denoted as "*ind* " = 1, otherwise 0.

$$ind = \begin{cases} 1, if\ \max(tp_1,...tp_n) - \max(tp_1,...tp_{(n-1)}) > 0 \\ 0, otherwise \end{cases}$$

$$ind = \begin{cases} 1, if\ \max(np_1,...np_n) - \max(np_1,...np_{(n-1)}) > 0 \\ 0, otherwise \end{cases}.$$

"n" is the times of tests taking. n=2, 3… 10.

The mean value of "*ind* " indicates the probability of improvement.

Figure 8 shows the comparison of the probability of improvement in total points among five NP subgroups. The combination of the first six bars represents the first NP subgroup that got normalized points belonging to the interval [0, 4]. Similarly, the other four combinations of the six bars represent the other NP groups respectively. There are six bars in every NP group, which represent six repeating groups and indicate the effects of test repeating 1 to 6 times. Probability of improvement increases as the number of gained point increases, which indicates that those test takers who got higher points have higher probability of improvement. Focus on each NP group, it can be seen that the probability of improvement decreases as times of repeat increases.
(The analysis is just based on the first six repeating groups due to the insufficiency of other data.)

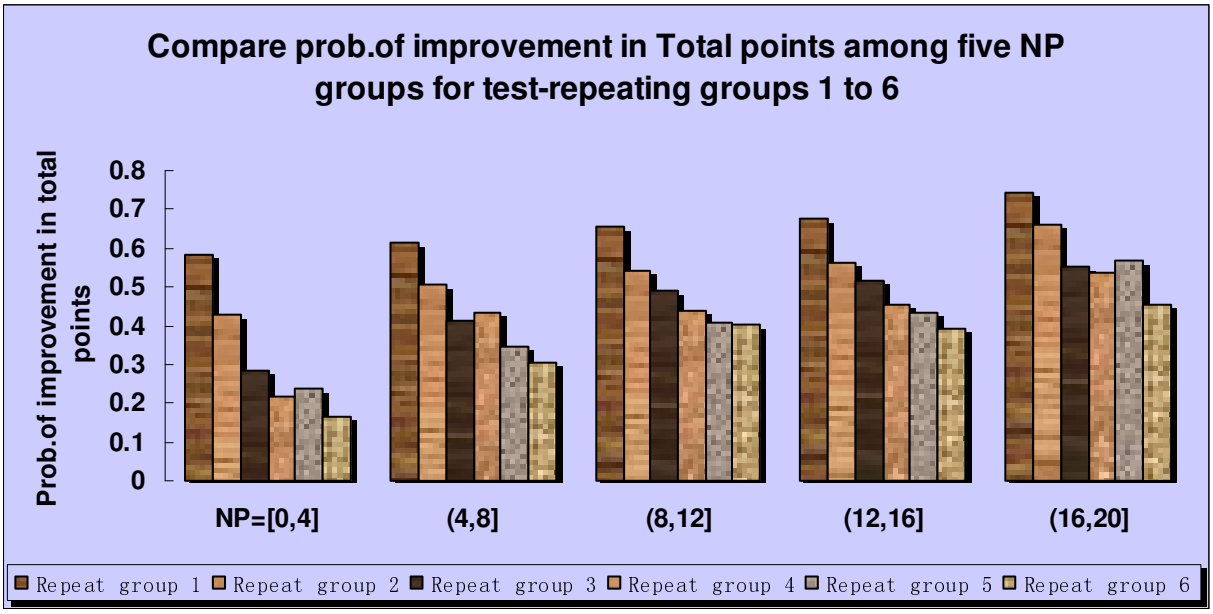

Figure 8: Probability of improvement in 5 NP subgroups for 6 different repeating groups

3.2.2 PROBABILITY OF IMPROVEMENT IN TOTAL POINT OBTAINED IN SUMULATION

As mentioned before, SweSAT contains 122 multiple-choice items with one item marked as 1 point if answering correct otherwise marked as 0. The sum of the points of the 122 items is the total point. Assume that the probability of answering correct on each item is P; however different test taker can have different probability of answering correct (different value of P). Therefore, the total point can be regarded as a binomial variable with 122 being the n trials and p the probability of success. In addition the analysis (probability of improvement in total point) in empirical data is simulated with fixed probability (0.2, 0.4, 0.6, 0.8, and 0.9). The formula is that TP ~Binomial (122, P), where P=0.2, 0.4, 0.6, 0.8 and 0.9

Figure 9 displays the probability of improvement in total points among five different probabilities of answering correct based on repeating groups 1 to 9. The five probabilities of answering correct should reveal different levels of test result, which are consistent with the roles of the five NP groups.

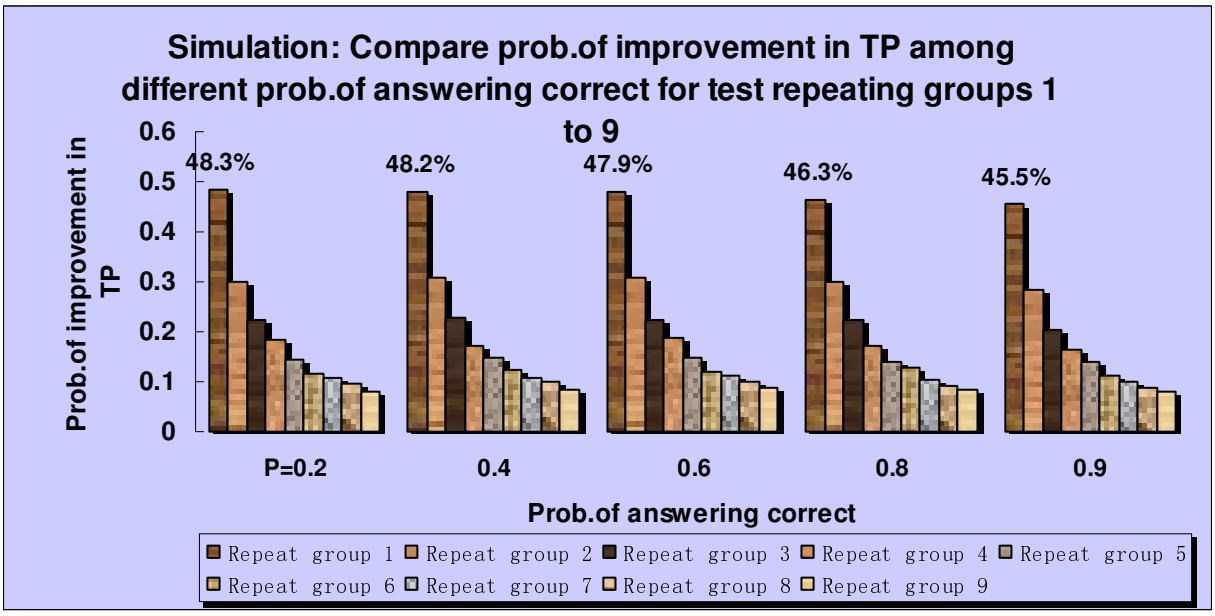

Figure 9: Probability of improvement in total points by repeating from 1 to 9 times for different probabilities of answering correct

For each NP group, real data and simulation get the same trend that is the probability of improvement will decrease as the times of repeat increases for each level of test result. On the other hand, the probability of improvement on each repeat occasion in real data is larger than that obtained in simulation, which indicates that there is learning effect for the test takers who repeated the test. So in each level of the test result, the probability of improvement declines more slowly than that in empirical result as times of repeating increases. The theoretical probability of improvement for the first repeating should be 0.5; however, the empirical data is less than that value. The half probability of improvement is just for one test taker; while the simulation result is obtained with 10000 sample size. It can not guarantee that the mean value of probability of improvement is 0.5 in simulation. Because each total point is regarded as random binomial variable; the difference between the final total point and the maximum of previous one is also random. So it is possible that the appearance of improvement is less than that of unimproved situation.

In addition, the probability of improvement in real data is increasing as NP level increases (Figure 8) in comparison with simulation where the probability of improvement is almost the same for different probability of answering correct. The conclusion will be that the learning had more effect on those test takers who gained higher points than on those who gained lower points. However, there may be interactivity between the probability of improvement and normalized point: higher probability of improvement also improves normalized point. The increasing trend in each kind of the bar among different NP groups shown in Figure 8 may be due to this reason.

Moreover, the comparison of probability of improvement in the first test repeating between good and low-point test takers with the first normalized point as the standard of NP groups division is done in Table 3. It can be seen that the lower point test takers have larger probability of improvement than higher point ones. However, it is still true that on average high point test takers have larger probability of improvement, which coincides with the preference of test repeating.

Table 3: Probability of improvement in normalized point

	[0,4]	**(4,8]**	**(8,12]**	**(12,16]**	**(16,20]**
1st repeat	67%	59%	55%	50%	46%
Mean prob.of improvement	38%	42%	41%	41%	39%

3.3 EFFECTS OF REPEATING ON THE PROBABILITY OF IMPROVEMENT IN FIVE SUBTESTS, TOTAL AND NORMALIZED POINTS

3.3.1 PROBABILITY OF IMPROVEMENT IN SUBTESTS, TP AND NP

Figure 10 considers the comparison of the probability of improvement in test results (five subtests, total and normalized points) among six repeating groups. Horizontal axis represents different repeating groups and each color of the bar represents one kind of subtest. Generally, as time of repeating increases, probability of improvement decreases in all tests results. Probability of improving in WORD tests is higher than in the other subtests and as the time of repeating increases this rule becomes obvious. In addition Table 4 in *Appendix B* shows the values of chance of improving in five subtests based on repeating groups 1 to 6. “To explain the score gained for the WORD test by referring to complexity of the item format is less plausible since the WORD subtest has a relatively simple format. It is also an item format that is familiar from school and vocabulary is also highly related to education. Therefore, the most plausible explanation is that the gain is an effect of schooling, i.e., the comprehension of words and

concepts increases during the period of schooling and this period coincides with the period of repeated test taking." [5]

It can be seen in Figure 11 that for test takers who repeated the test once, the probabilities of improvement in total and normalized points on average are 64% and 57% respectively. However, the probability of improvement in NP is less than that in TP on each repeating occasion, which means that the NP is not easy to be improved. Because if the large improvement happen in total points which may be due to the easiness of the examination, it will be scaled down by transforming TP into NP. As mentioned in Introduction section, NP is used as a consistent standard to make test results comparable with each other.

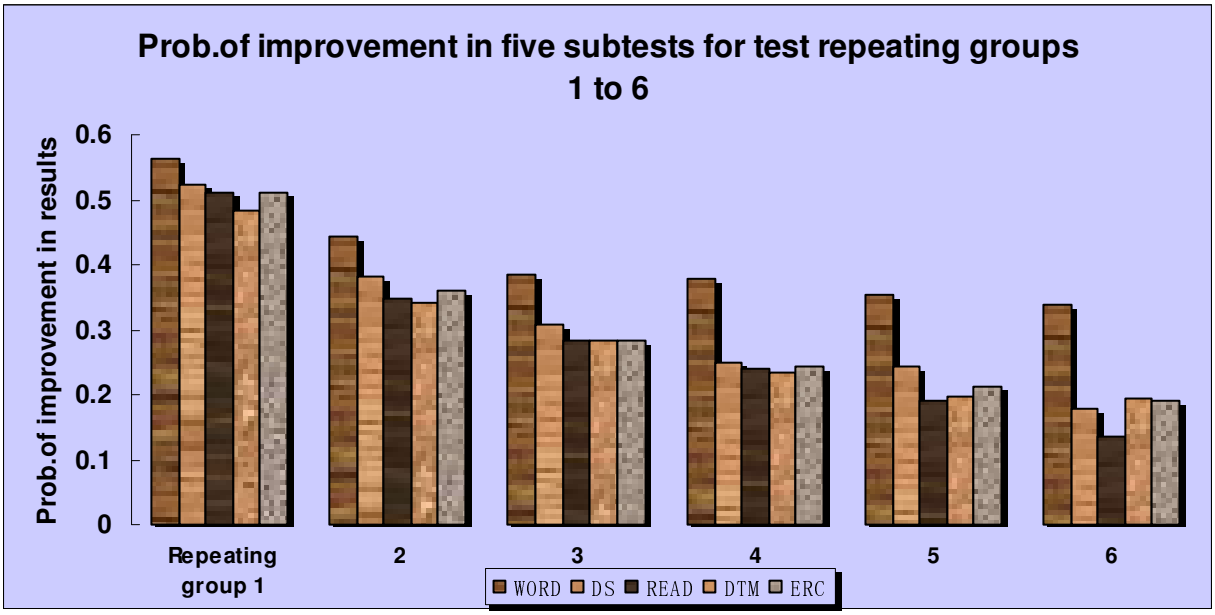

Figure 10: Probability of improvement in subtests for repeating groups 1 to 6

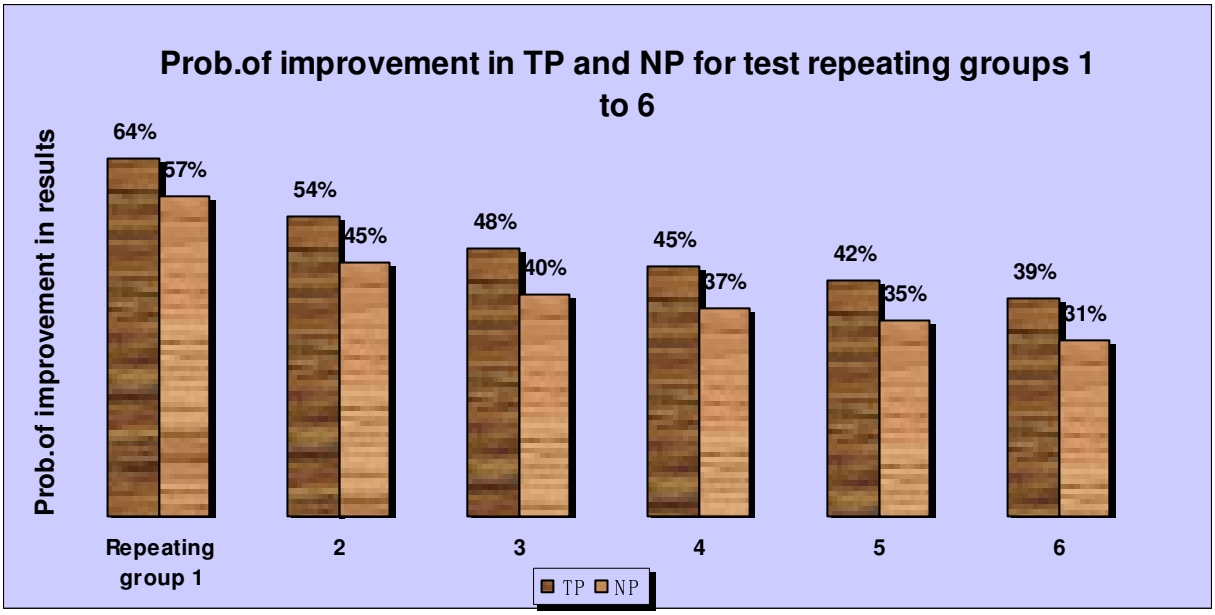

Figure 11: Probability of improvement in total and normalized points for repeating groups 1 to 6

3.3.2 COMPARE THE PROBABILITY OF IMPROVEMENT IN EACH TIME OF REPEATING AND IN EACH REPEATING GROUP

Table 5 shows the probability of improvement in total and normalized point by repeating the test several times. The figures are calculated by analyzing all the effects of repeating. i.e. Test takers who repeated test nine times, their repetition effect from first until ninth time should be included in each calculation of the column of table 5. For test takers who repeated test eight times, their effect of test repeating first to eighth times also should be analyzed in the first to eighth columns of table 5 and so on.

Table 5: Probability of improvement in total and normalized points: concentrate on the first to ninth repeating of all the test takers

Repeat effect	1st	2nd	3rd	4th	5th	6th	7th	8th	9th
TP	65%	55%	49%	44%	43%	35%	39%	43%	27%
NP	58%	46%	39%	35%	34%	29%	25%	25%	18%

Table 6 displays the probability of improvement in the final total and normalized points in each repeating group. i.e. for test-repeating group 4, considering that if the last result of each test-taker is larger than the maximum point of the previous 4 total points; "*ind* " is evaluated and its mean value is the chance of improvement in the final result through just repeating the test 4 times.

Table 6: Probability of improvement in total and normalized point in each test repeating group

Repeating groups	1	2	3	4	5	6	7	8	9
TP	64%	54%	48%	45%	42%	39%	38%	42%	27%
NP	57%	45%	40%	37%	35%	31%	27%	30%	18%

The output in Table 6 is consistent with that in Table 5, so the analyses in repeating groups 1 to 9 are enough to show the effect of test repeating 1 to 9 times.

3.4. THE EFFECT OF REPEATING ON THE CHANGE IN TOTAL AND NORMALIZED POINTS

In order to see the effect of test-repeating 1 to 6 times on total and normalized points, "change" is defined as the difference between the maximum value of all tests results (n times) and that of the previous tests (n-1 times) for each test taker. The formula is as follows: for each test repeating group

$$imp = \max(tp_1,...tp_n) - \max(tp_1,...tp_{(n-1)})$$
$$change = mean(imp)$$
$$imp = \max(np_1,...np_n) - \max(np_1,...np_{(n-1)})$$
$$change = mean(imp)$$

, where "n" is the times of tests taking". n=2, 3 … 7.

The mean value of the difference is obtained as the effect of that test repeating on the "change" of the test results.

The deep color of the bar in Figure 12 represents the change in total point while the light color represents the codes of change in normalized point. Vertical axis shows six test repeating groups.

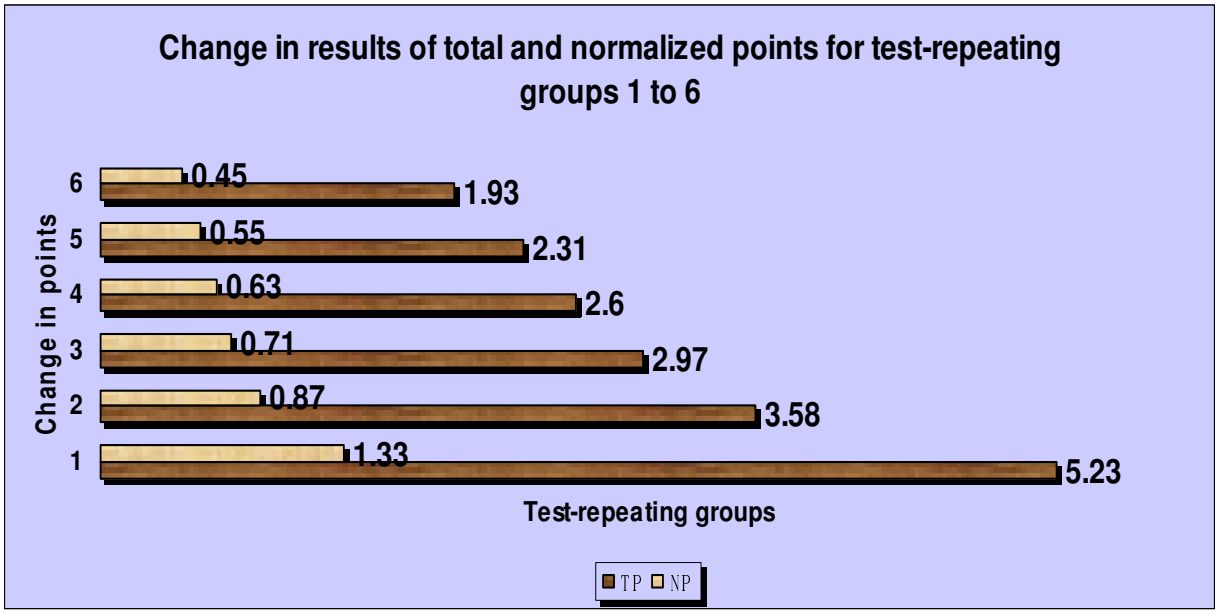

Figure 12: Change in results of total and normal points for different repeating groups

It can be seen that the largest improvement occurs in the first test repeating, about 5.23 and 0.13 points of improvements in total and normalized points respectively. The positive change in results is a declining function on the times of repeating. However, even for the group who repeated the test 6 times, they still can have 0.045 points of improvement averagely compared to the maximum normal points they got in the last results.

3.5. COMPARISON OF THE PROBABILITY OF IMPROVEMENT IN NP FOR SUBGROUPS (GENDER/AGE) BASED ON EACH REPETING GROUPS

Figure 13 shows the comparison between men and women based on each test repeating group in terms of probability of improvement in normal points. Men have larger probability of improving in each repeating group comparing with women. As mentioned in section 3.2.1, the test takers

who got higher points have larger probability of improvement. It is also true in the comparison of the normalized point between men and women that men have higher points. (Figure 6)

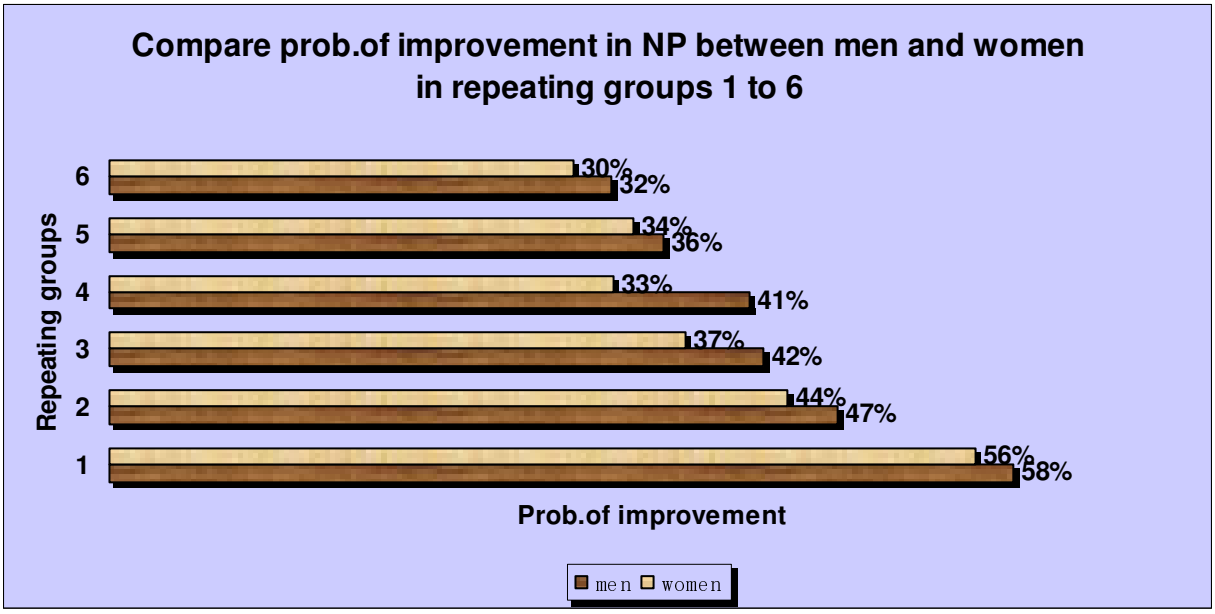

Figure 13: Comparison of the probability of improvement in NP by repeating the test from 1 to 6 times in different gender group

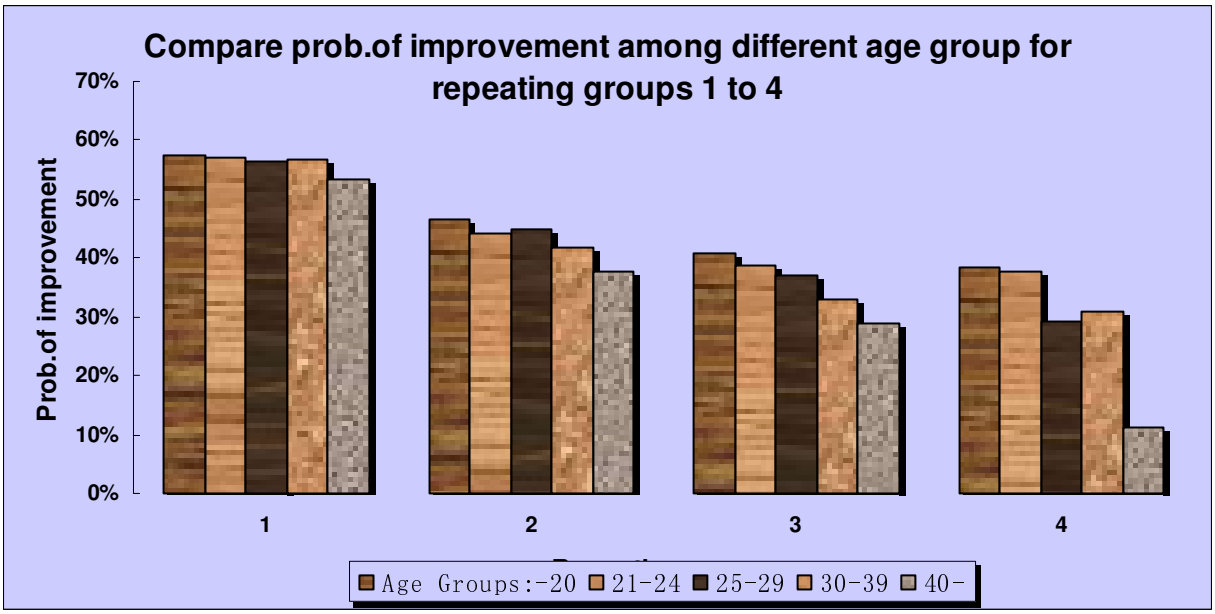

Figure 14: Comparison of the probabilities of improvement in normal points by repeating the test for different age groups

Comparison of probability of improvement among different age group is shown in Figure 14. The probability of improvement is decreasing when the age of tests-taker increases on each test

repeating occasion. The big difference in the 4th repeating group may due to the relatively fewer test takers in age group 5.

Table 7 & 8 in *Appendix B* show the whole information (including subtest and total point) about the probability of improvement for gender groups and age groups. The conclusion is that men have larger probability of improvement in WORD, DS and READ. The differences of men and women in DTM and ERC are not large in term of probability of improvement. The difference of probability of improvement in subtest among age groups is not big in the first test repeating. However, as times of repeating increases, the probability of improvement for old test takers decreases more quickly than for young ones. In terms of total point, men have larger probability of improvement than women; and young test takers have larger chance of improvement than old ones.

4. DISCUSSION

The summarized conclusions:
There are generally 42045 applicants who took the SweSAT offered in April, 31636 people who took in October. On average 37% of the test takers are repeaters on each test occasion during the years 2004 to 2006. There are more women (55%) than men (45%) who took the test. However, men have higher test results (total and normalized point) than women in each test taking group. Around 44% of the test-takers are under 20 years old. The difference in normalized points obtained between five age groups is small.

Test takers who have gained better score would like to repeat the test and they have higher probability of improvement than those who got low points. This could happen also because that high probability of improvement has positive effect on normalized point. However, in general test takers with good point have larger probability of improvement and this fact is conformed by the preference of test takers with different levels of result. Comparison of the probability of improvement obtained in real data and R simulation shows that test takers can increase their result by learning. And test takers who got high points have larger ability of improvement by learning.

Both men and women can improve their scores through repeating the test. The probability of improvement decreases as the times of repeating increases. Men have higher probability of improvement than women in test results (total and normalized point).

On average, elder test takers got better scores than young ones in the first test. Generally younger test takers have a comparatively higher probability of improvement than older ones, though the difference is not large.

“The summarized results of the effects of repeated test taking indicates that the highest gain of repeated test taking is from the first to the second test occasion due to test wise ness (TW) (test takers will then be familiar with the requirements of the test and the test situation).”[4]

In conclusion: improvement in WORD subtest is largest compared with other subtests; Improvement (probability of improvement & positive change in test results) are made by learning on each repeating occasion; in addition men, younger, and test takers with high points have relatively larger chance of improvement by repeating SweSAT, although the improvement is a decreasing function of the test repeating.

REFERENCES

[1] The Swedish Scholastic Assessment Test - SweSAT, Department of Educational Measurement, Umea University, http://www.umu.se/edmeas/hprov/index_eng.html

[2] Hogskoleverket.(1977a).Examination at universities and in higher education – from the student's point of view

[3] Christina Stage, Entrance to higher education in Sweden, EM No 51, 2004

[4] Birgitta Tornkvist, Widar Henriksson, Validity issues concerning repeated test taking of the SweSAT, EM No 56, 2006

[5] Birgitta Tornkvist, Widar Henriksson, SweSAT REPEAT, EM No 46, 2004

[6] Christina Stage, EXPERIENCES WITH THE SWEDISH SCHOLASTIC APTITUDE TEST (EM No 15)

[7] Henriksson,W.(1983).The problem of repeated test taking and the SweSAT. (EM,No 5). Division of educational Measurement, Department of Education, University of Umea

[8] Christina Stage, Gunilla Ögren**.** The Swedish Scholastic Assessment Test (SweSAT). **(**EM No 49**).** Development, Results and Experiences.

[9] Widar Henriksson (1995) REPEATED TEST TAKING AND THE SweSAT (EM No 13.) Division of educational Measurement, Department of Education, University of Umea

[10]Peter Dalgaard (2002), INTRODUCTORY STATISTICS WITH R, Department of Biostatistics, University of Copenhagen

APPENDICES

APPENDIX A: FIGURES

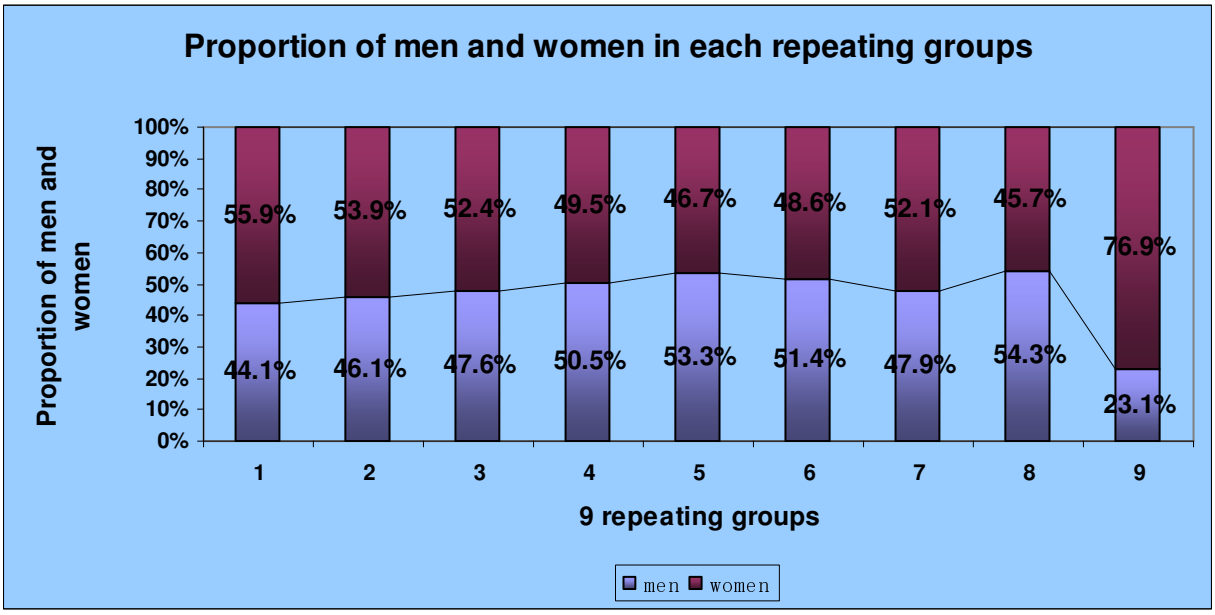

Figure 1: Proportion of men and women in each repeating groups

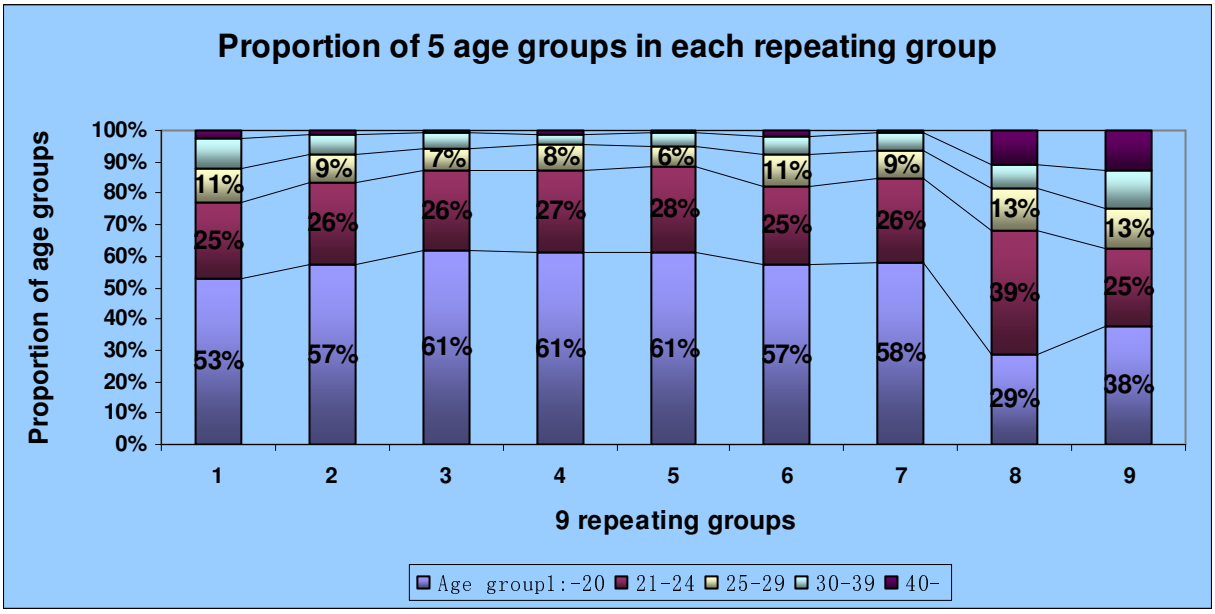

Figure 2: Proportion of five age groups in each repeating groups

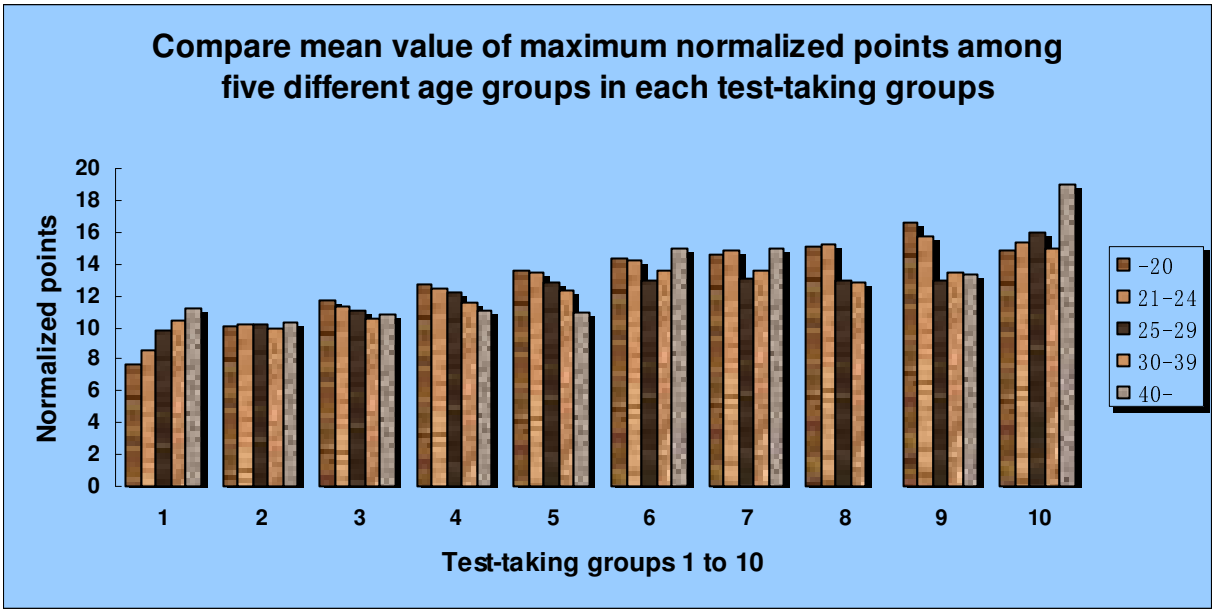

Figure 7: Compare average value of maximum normalized points among five age groups in each test-taking group

APPENDIX B: TABLES

Table 1: Number of tests-takers, repeaters and their percentage in each test offered during years from 2002 to 2006

Each tests offered	Number of test-takers	Number of repeaters	Proportion of repeaters
02A	40361		
02B	28510		
03A	40428		
03B	31638		
04A	42017	14561	34.66%
04B	35232	12402	35.20%
05A	46105	17411	37.76%
05B	33109	12417	37.50%
06A	41313	16414	39.73%
06B	29689	11797	39.74%
Average	36840	14167	37.43%

Table 2: Distribution of normalized points (NP) over repeated tests (both in number of people and proportion)

Repeating group / NP	1	2	3	4	5	6	7	8	9	Total	Proportion
[0,4]	6925	1370	311	78	21	6	3	0	0	8714	11.78%
(4,8]	13997	3803	1061	278	81	23	11	2	0	19256	26.03%
(8,12]	15888	5518	1898	603	227	94	35	9	1	24273	*32.81%*
(12,16]	10225	3990	1712	729	345	143	51	19	9	17223	23.28%
(16,20]	2680	1057	487	177	65	22	17	3	1	4509	6.10%
Total	49715	15738	5469	1865	739	288	117	33	11	73975	100%
Proportion	*67.21 %*	21.27 %	7.39%	2.52%	1.00%	0.39%	0.16%	0.04%	0.01%	100%	

Table 4: Probability of improvement in normalized point in five subtests in repeating group 1 to 6

Repeat group	WORD	DS	READ	DTM	ERC
1	56.40%	52.40%	51.00%	48.20%	51.00%
2	44.30%	38.10%	34.70%	34.20%	35.90%
3	38.60%	30.80%	28.40%	28.30%	28.20%
4	38.00%	24.80%	23.90%	23.40%	24.30%
5	35.50%	24.20%	19.20%	19.60%	21.20%
6	34.00%	17.70%	13.50%	19.40%	19.10%

Table 7: Compare the probability of improvement for men and women by repeating tests 1 to 9 times

		1	2	3	4	5	6	7	8	9
WORD	**men**	58.2%	46.0%	41.4%	40.3%	36.3%	38.5%	46.4%	27.8%	50.0%
	women	55.0%	42.9%	36.0%	35.5%	34.5%	29.3%	26.2%	46.7%	11.1%
DS	**men**	53.8%	39.6%	31.6%	27.8%	25.1%	18.9%	28.6%	16.7%	50.0%
	women	51.2%	36.8%	30.1%	21.7%	23.2%	16.4%	14.8%	20.0%	0.0%
READ	**men**	51.6%	35.2%	29.1%	24.3%	19.3%	14.9%	10.7%	5.6%	0.0%
	women	50.6%	34.3%	27.8%	23.4%	19.1%	12.1%	16.4%	20.0%	0.0%
DTM	**men**	48.7%	34.6%	28.6%	23.2%	19.8%	16.2%	16.1%	11.1%	0.0%
	women	47.8%	33.9%	28.0%	23.5%	19.4%	22.9%	6.6%	33.3%	11.1%
ERC	**men**	50.9%	36.1%	28.9%	25.4%	22.1%	18.2%	10.7%	16.7%	50.0%
	women	51.0%	35.8%	27.6%	23.3%	20.3%	20.0%	18.0%	13.3%	11.1%
TP	**men**	65.4%	55.7%	50.9%	49.7%	44.4%	41.2%	42.9%	38.9%	50.0%
	women	63.1%	51.8%	44.8%	39.2%	39.7%	36.4%	32.8%	46.7%	22.2%
NP	**men**	58.4%	47.0%	42.2%	41.3%	35.8%	32.4%	28.6%	27.8%	50.0%
	women	55.9%	43.8%	37.2%	32.6%	33.9%	30.0%	26.2%	33.3%	11.1%

Table 8: Compare the probability of improvement for different age groups by repeating tests 1 to 9 times

		1	2	3	4	5	6	7	8	9
WORD	**-20**	56.8%	44.9%	40.0%	39.8%	37.4%	35.9%	40.0%	40.0%	0.0%
	21-24	57.2%	45.9%	38.0%	37.1%	32.8%	31.0%	29.0%	28.6%	33.3%
	25-29	55.0%	42.7%	34.0%	33.3%	34.9%	33.3%	30.0%	75.0%	100.0%
	30-39	54.4%	37.3%	33.1%	29.4%	25.0%	18.8%	33.3%	50.0%	0.0%
	40-	53.3%	32.1%	21.1%	16.7%	40.0%	75.0%		0.0%	0.0%
DS	**-20**	52.8%	38.5%	31.3%	25.8%	24.2%	15.6%	25.7%	0.0%	0.0%
	21-24	52.3%	37.3%	30.6%	26.5%	24.5%	19.7%	16.1%	28.6%	33.3%
	25-29	52.0%	37.9%	28.7%	14.6%	27.9%	23.3%	20.0%	25.0%	0.0%
	30-39	50.9%	37.8%	29.7%	19.1%	12.5%	25.0%	0.0%	0.0%	0.0%
	40-	50.3%	39.3%	21.1%	16.7%	60.0%	0.0%		33.3%	0.0%
READ	**-20**	51.4%	35.4%	29.6%	24.6%	19.8%	16.8%	17.1%	10.0%	0.0%
	21-24	50.3%	34.4%	27.7%	22.1%	19.1%	11.3%	12.9%	21.4%	0.0%
	25-29	50.9%	34.5%	24.4%	28.5%	16.3%	3.3%	0.0%	0.0%	0.0%
	30-39	51.2%	30.9%	24.8%	14.7%	18.8%	12.5%	0.0%	0.0%	0.0%
	40-	49.2%	29.1%	15.8%	22.2%	0.0%	0.0%		0.0%	0.0%
DTM	**-20**	48.1%	34.7%	28.9%	23.9%	20.7%	17.4%	11.4%	10.0%	20.0%
	21-24	47.7%	34.1%	27.2%	25.9%	16.7%	29.6%	9.7%	21.4%	0.0%
	25-29	49.1%	33.6%	27.7%	16.7%	20.9%	13.3%	20.0%	0.0%	0.0%
	30-39	49.1%	31.1%	27.8%	10.3%	21.9%	12.5%	0.0%	50.0%	0.0%
	40-	46.6%	33.7%	18.4%	22.2%	20.0%	0.0%		66.7%	0.0%
ERC	**-20**	51.4%	37.1%	30.0%	27.4%	21.3%	20.4%	22.9%	10.0%	40.0%
	21-24	50.3%	34.6%	26.6%	19.9%	22.1%	16.9%	3.2%	28.6%	0.0%
	25-29	49.9%	34.1%	22.6%	18.8%	16.3%	13.3%	0.0%	0.0%	0.0%
	30-39	51.1%	33.4%	24.1%	22.1%	21.9%	31.3%	0.0%	0.0%	0.0%
	40-	52.7%	31.6%	18.4%	5.6%	20.0%	0.0%		0.0%	0.0%
TP	**-20**	64.6%	54.7%	48.9%	46.2%	43.1%	37.7%	41.4%	40.0%	20.0%
	21-24	63.7%	52.9%	47.7%	45.4%	42.2%	39.4%	32.3%	42.9%	33.3%
	25-29	63.2%	52.5%	42.6%	36.1%	39.5%	36.7%	50.0%	50.0%	100.0%
	30-39	64.1%	49.4%	41.7%	35.3%	34.4%	50.0%	0.0%	50.0%	0.0%
	40-	61.8%	46.9%	36.8%	16.7%	40.0%	50.0%		33.3%	0.0%
NP	**-20**	57.4%	46.4%	40.7%	38.4%	35.6%	28.7%	32.9%	40.0%	20.0%
	21-24	57.0%	44.2%	38.9%	37.8%	34.3%	35.2%	22.6%	28.6%	33.3%
	25-29	56.3%	44.8%	37.1%	29.2%	34.9%	36.7%	20.0%	25.0%	0.0%
	30-39	56.8%	41.7%	33.1%	30.9%	31.3%	31.3%	0.0%	50.0%	0.0%
	40-	53.3%	37.8%	28.9%	11.1%	20.0%	25.0%		0.0%	0.0%