Notes from the Sixth International SweSAT Conference

San Diego April 12, 1998.

Christina Stage

Introduction

The International Scientific Advisory Board to the SweSAT program, which was set up in 1992, met for the sixth time in San Diego, USA in April, 1998.

In this report a reproduction of the presentations and a summary of the discussions at this sixth meeting are presented. A list of participants is enclosed as Appendix 1.

Welcome and opening address

Christina Stage

Christina Stage opened the conference by welcoming all the members and especially the new member Vice-President, Assessment Division at ETS, Dr. Linda Cook. Warm thanks were extended to professor Ronald Hambleton, who had kindly arranged the meeting accommodations.

A minute of silence was held in remembrance of professor emeritus Sten Henrysson, the founder of SweSAT, who died on March 17th.

The SweSAT Program since May 1997

No new features had been introduced in the test during the last year. The test and pretesting procedures which were introduced in 1996 are still used.

The most important issue for the project during the past year has been to follow up and evaluate the new pretesting procedure. Since the present procedure is only on a trial basis, a decision about future procedure will be based on this evaluation. One important experience, so far, is that the development of new pretest material is very costly. Since it is important that the testtakers should not be able to distinguish the pretest booklets from the regular test booklets, the pretest material must be reviewed and compiled much more carefully than before. The results of the pretesting, however, are much more stable, than they were before and there is no doubt that this method is better than the old one. Some new issues have come to attention, which were not noted when a more crude pretesting model was used. Especially for the

READ subtest it is of great importance whether a text is placed as number one or number five. For some of the other subtests it seems to be of some importance whether an item is presented in the beginning or at the end of the subtest.

The number of testtakers was higher than ever in 1997 but still the project has more financial problems than before, which primarily affects the research activities.

The Swedish *Advisory Board for SweSAT and other Admission Tests* constituted by the National Agency[1] for Higher Education has met once a month during the last year and has so far mainly discussed ”other admission tests” and admission rules.

Overview of SAT I and SAT II

Linda Cook

Overview of SAT I and II
Presentation for the International
Advisory Board for the SweSAT
April 12, 1998

Good Morning. I’m very pleased that Christina has invited me to be a member of the SweSAT Advisory Board and also that she has given me the opportunity to talk to all of you today about the SAT that ETS administers for the College Board. Christina suggested that I focus my discussion on the kinds of abilities measured by the SAT, the essay that is given as part of the SAT II administration, and our pretesting procedures. I most certainly plan to cover all three of these topics. But since this is my first opportunity to speak to all of you about the SAT, and since Christina has graciously given me an hour for the discussion, I’d like to provide some background about the SAT as well as covering the three topics Christina suggested.

I thought I would cover the following topics today.

Overview of Presentation

- Background of the SAT
- Overview of SAT I and SAT II
- The Test Development Process
- Test Creation Reengineering

[1] The National Agency for Higher Education was constituted in 1995 and has since then the governmental responsibility for SweSAT.

I would like to begin by talking briefly about the history and the background of the SAT, than I would like to spend time giving you an overview of both the SAT I and SAT II. I'll talk about what the tests measure and how they are formatted. Next I'd like to talk about the test development process, including the pretesting process, and finally I'd like to spend a short amount of time talking about our test creation reengineering project. We are currently in the middle of reengineering our entire test creation process and I thought you might like to hear about some of the changes we are making.

So, first I'll say a few words about the history of the SAT.

History of the SAT

- First College Board Admissions Tests Administered in 1901
- First SAT Administered in 1926
- SAT Scale Set in 1941
- New SAT Introduced in 1994
- SAT Rescaled in 1995

SAT I and SAT II Tests

- Purpose
  - Common standard for student performance
  - Test scores recommended for use in conjunction with other information
- SAT I Reasoning Test
  - Verbal and Math abilities
- SAT II Subject Tests
  - 24 tests measuring academic achievement

The original purpose of the SAT was to provide a common standard for student performance. This common standard is necessary because in the United States, high school courses and also grading practices differ greatly from school to school. The College Board and ETS recommend that colleges and universities use as much information as possible when making admission decisions—we recommend that schools use test scores, course grades, and other relevant information like extra curricular and community activities.

Today, the SAT testing program consists of the SAT I reasoning test which measure developed verbal and mathematical abilities and the SAT II.

Subject tests, which are tests designed to measured academic achievement. There are currently 24 different subject matter tests. What I'd like to do next is spend some time talking about each of these testing programs.

Test Administrations

- SAT I
  - Seven domestic administrations
  - Six international administrations
  - 4,500 test administration sites
  - Annual volume of approximately 2 million candidates
- SAT II
  - Six domestic and six international administrations
  - Annual volume of approximately 400,000 candidates

The SAT I is administered 7 times a year domestically and 6 times a year internationally. We currently have 4,500 test administation sites in this country, most of these sites are in high schools, but a few are in colleges, regional offices and Sylvan Centers. The annual testing volume for the SAT is about 2 million students.

The SAT II Subject tests are administered both domestically and internationally six times a year. They are administered in the same sites as the SAT I Test. The annual volume for the SAT II Subject tests is approximately 400 000 students.

Before I describe the current SAT I and II tests, I'd like to say a few words about the changes that were introduced to the SAT in 1994.

The New SAT

- Introduced Spring 1994
- Began With Special Task Force in 1986
- Primary Reasons For Change
  - Increase educational relevance of tests
  - Ensure tests reflect current trends in curriculum and education
  - Incorporate advances in test design
- Additional Reasons For Change
  - Reduce gender differences on verbal test
  - Reduce potential for coaching
  - Make test preparation more relevant for schoolwork
  - Improve and better integrate SAT II tests

The first New SAT was administered in the spring of 1994. The changes to the test that were introduced were the result of an extensive research and development project that began when a special Task Force was assembled in 1986. There were a number of reasons why the College Board and ETS decided it was time to revise the SAT. Among the primary reasons was a desire to increase the educational relevance of the test content and format; to ensure that the test reflected current educational and curricular trends; and to ensure that the new test incorporated advances in test design that were available because of new technology such as item response theory. In addition, we were concerned about reducing gender differences on the verbal test that were continuing to increase. We wanted to reduce the potential for coaching the test; to make test prep more related to school work; to decrease test speededness and to improve and better integrate the SAT II tests into the testing program.

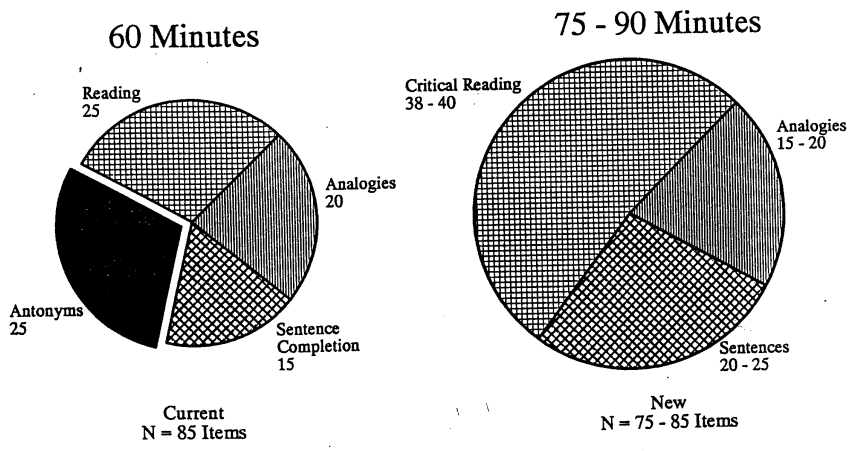
I'll start by describing the changes made to the verbal test.

Changes to SAT Verbal

- Greater Emphasis on Critical Reading and Reasoning
- Longer Reading Passages
- More Accessible Reading Material
- Double Passage Added
- Passage-based Vocabulary Questions and Vocabulary in Context

The major changes that were made to the verbal SAT I test were changes to the reading measure. The goal was to place a greater emphasis on critical reading and reasoning. Longer reading passages were added, the reading material was made more accessible and engaging, we added a double passage with two points of view on the same subject, we added introductory and contextual information for the reading passages, and we added passage based vocabulary questions that tested vocabulary in context.

CONTENT & FORMAT CHANGES TO THE VERBAL TEST



The slide you are looking at now summarizes the content and format changes that we made to the verbal test. A major change was increasing the total testing time from 60 to 75 minutes, while decreasing the total number of items from 85 to 78. The number of critical reading items was increased from 25 to 40, antonyms were dropped and the number of analogies and sentence completions was changed to 19 apiece.

We also substantially changed the SAT math test.

Changes to SAT Mathematical

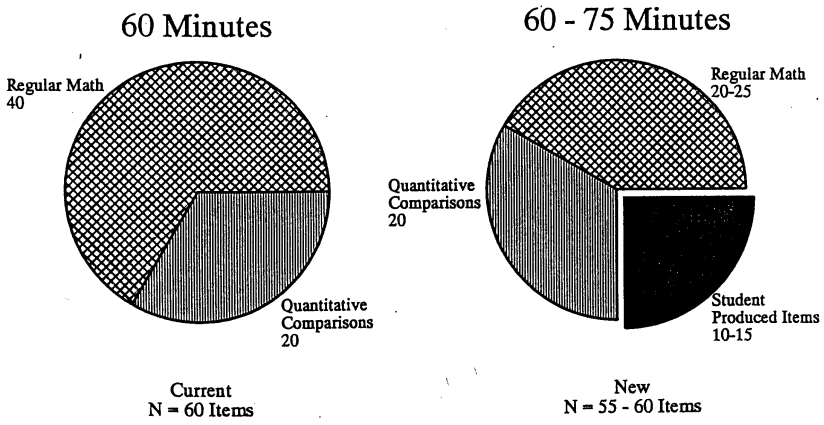
- Addition of Non-multiple-choice Questions
- Calculator Use Permitted
- Continued Emphasis on Problem Solving
- Increased Emphasis on Data Interpretation and Applied Mathematics

Changes made to the SAT I mathematical test included the addition of non multiple-choice questions that require students to produce and

grid-in their own answers to some questions; we permitted the use of a calculator on the test; we continued the tests's emphasis on problem solving in arithmetic, algebra, and geometry; we increased the test's emphasis on data interpretation and applied mathematics questions.

CONTENT & FORMAT CHANGES TO THE MATHEMATICAL TEST

Mathematical



The slide you are looking at now summarizes the changes made to the math test. The total testing time for the math test was changed from 60 to 75 minutes, the total number of items was kept at sixty. You can see that the number of quantitative comparison items remained the same, we reduced the number of regular math items to make room for the items with the student produced responses.

Changes to SAT II Subject Tests

- New Tests in Asian Languages
- Listening Component Added to Foreign Language Tests
- English As a Second Language Proficiency Test
- Advanced Math Tests Designed for Calculator Use

The major changes to the SAT II Subject tests that were made in 1994 were the addition of new tests in Asian languages, the addition of a listening component to foreign language tests, the addition of a new English as a Second Language Proficiency Test, the modification of the Mathematics Level I and II tests to require the use of calculators, and the replacement of the English Composition Test with Essay with the new SAT II Writing test. I'll say more about the SAT II Writing test in a few minutes.

What I'd like to do next is describe the current SAT I tests in more detail.

SAT Verbal Test

- Measures Verbal Reasoning
- Two 30-Minute and One 15-Minute Section
- Three Types of Questions
- Analogies
  - Sentence Completions
  - Critical Reasoning

The SAT verbal test is a multiple-choice test of developed verbal reasoning ability. The test consists of two 30 minute sections and one 15 minute section. Three types of questions are used on the SAT I, Analogies, Sentence Completions, and Critical Reading Questions.

Next I'd like to show you an example of each item type.

Critical Reading Questions

- 40 Critical-Reading Questions
- Four Passages (400-850 words each)
- Measure
- Vocabulary in context
- Literal comprehension
- Extended reasoning
- Content Areas
- Humanities
- Social Sciences
- Science
- Narratives

The SAT verbal test contains 40 critical reading questions. These questions measure the student's ability to read and think carefully about the reading passages included in this is the use of vocabulary in context, literal comprehension and extended reasoning. Each test contains four passages of varying length (400-850 words). Content areas include Humanities, social Sciences, Sciences and Narratives. At

least one set of passages will be a double passage, a passage that will have two points of view. The passage I've included in your hand out is an example of a double passage taken from a recent SAT (Appendix 2).

The passages were written by James Baldwin and Malcom X and each passage describes significant influences during the authors' formative years. The questions following the passages focus on the specific passages but also ask the student to compare and contrast the passages.

I'm sure you are all familiar with the analogy items that appear on the SAT. Your hand out contains a sample of these items that was taken from a recent SAT (Appendix 3).

Analogies
- 19 Analogy Questions
- Measure
  - Word meaning
  - Ability to see relationship in pair of words
  - Ability to recognize similar or parallel relationships

Each SAT verbal test contains 19 Analogy items. These items measure knowledge of the meaning of words, the ability to see a relationship in a pair of words, the ability to recognize similar or parallel relationships.

I've also included an examples of Sentence Completions items in your handout (Appendix 4).

Sentence Completions
- 19 Sentence Completion Questions
- Measure
  - Word meaning
  - Ability to understand how different parts of the sentence logically fit together

There are also 19 sentence completions items in each SAT verbal test. Sentence Completions items are intended to measure word meaning, but also the ability to understand how different parts of the sentence logically fit together.

I'd like to talk about the SAT mathematics test next.

SAT Mathematics Test

- Two 30-Minute and one 15-Minute Section
- Question Types
  - Standard 5-multiple choice
  - Student-produced responses
  - Quantitative comparisons
- Content Areas
  - Algebra
  - Geometry
  - Arithmetic
  - Applied math/data interpretation

The SAT Mathematics Test has two 30 minutes sections and one 15 minutes section. The test contains three questions types; standard 5-choice multiple choice questions; 4 choice quantitative comparisons questions; and student produced response questions that provide no answer choice. The math questions measure data interpretation, applied math, arithmetic, algebra, and geometry.

I'm sure you are familiar with the Standard five choice item type, but you may not be familiar with either the quantitative comparison item type or the student produced response item type. I'd like to show you examples of each one of these questions types (Appendix 5).

The student produced response item presents a problem to a student, but does not provide a set of forced choices. The student is required to generate the answer to the problem and then to enter that answer into a grid that is machine scoreable.

The view graph you are looking at now shows the directions for gridding a response to a student produced response item—notice that the student can grid a response as a decimal or a fraction –also they are able to right or left justify an number when they are gridding a response.

Next, I'll show you an example of a quantitative comparisons item (Appendix 6).

Quantitative Comparisons items are four choice items measuring math ability in arithmetic, algebra, and geometry.

The view graph that you are looking at now shows an example of this item type. Notice that the student is asked to compare the quantities in column A and column B and to ascertain if the quantity in column A is greater, if the quantity in column B is greater, if the quantities are

equal, or if the relationship cannot be determined from the information given.

Next, I'll say a few words about the SAT II Subject Tests.

SAT II Subject Tests

- Writing
- Literature
- Foreign Languages
- History and Social Studies
- Mathematics
- Sciences

The SAT II subject tests are designed to measure knowledge and skills in a particular subject area and the students ability to apply that knowledge. The major categories that subject tests are given in are: writing, literature, foreign languages, history and social studies, mathematics, and sciences. All tests, except the writing test, are one hour multiple choice tests. The writing test consists of a 20 minute essay and a 40 minute multiple choice section.

I'd like to spend some time now talking about the Writing test.

SAT II Writing Test

- Administered Six Times a Year
- 60 Multiple-choice Questions
- One 20-minute Essay
- Measures
  - Ability to express ideas effectively
  - Recognition of faults in usage and structure
  - Use of language with sensitivity to meaning
- Item Types
  - Identifying sentence errors
  - Improving sentences
  - Improving paragraphs

The SAT II Writing test is administered six times a year as part of the Subject Test battery. The test contains 60 multiple choice questions that measure a students ability to express ideas effectively in standard written English, to recognize faults in usage and structure, and to use language with sensitivity to meaning. The test contains three types of multiple choice items, Identifying Sentence Errors, Improving Sentences, and Improving Paragraphs. The test also contains a 20 minute essay. Each essay is read and scored by two readers on a six point scale.

The two multiple choice item types, Improving Sentences, and Identifying Sentence Errors are quite straight forward—there are examples in the program publications that I will give to all of you, so I'm not going to go over those now. The Item types that I would like to spend some time discussing are the essay and the Improving Paragraphs item type. I'll talk about the essay first.

SAT II Essay Prompt

"I have experienced various things that have made me feel worthwhile, but I have never felt better than when ----------."

The slide you are looking at now shows an essay prompt from a recent administration of the SAT Writing test—Students are asked to write an essay completing the statement –they are asked to be sure to explain the reasons for their choice.

Students are told that they have twenty minutes to write an essay and that they must write on topic—they are cautioned that the essay tests how well they write, not how much –they are told that one well written paragraph is probably sufficient.

SAT II Improving Paragraphs

- Students Presented with Draft Essay
- Multiple-choice Questions Ask Students to Improve
  - Sentence structure
  - Choice of words
  - Paragraph organization
  - Paragraph development

Your hand out contains an example of the Improving Paragraphs item type (Appendix 7).

This item type provides an early draft of an essay. Students are asked to respond to multiple choice questions that will improve the draft by improving the sentence structure, the choice of words, or maybe even the entire organization and development of the essay.

I think you can see that what we are really trying to do with the SAT Writing test is to provide the student with the opportunity to show a progression of skills, from basic skills that require recognizing errors in sentences, to more sophisticated skills that show the students ability to recognize errors in paragraphs, to finally providing the student with an opportunity to demonstrate their writing skills through the essay.

I'd like to switch topics now and spend some time talking about the test development process for the SAT I and II tests.

We are right in the middle of reengineering our test development process, so the process I'm going to describe to you is the old process. What you will see is that it is very slow and labor intensive.

SAT Test Development

- Who Designs the SAT?
- Basic Constructs Introduced in 1929
- Most Recent Modification to SAT in 1994
- Our Approach to Measuring Verbal and Math Reasoning Must Continue to Evolve

We are right in the middle of reengineering our test development process. The SAT has not moved to the new process yet, so I'm going to describe the old process to you. What you will see is that it is slow and very labor intensive.

People often ask, "who designs the SAT?" The basic constructs measured by the test, Verbal and Math reasoning, were part of the SAT when it was first introduced in 1929 and they do have their roots in early intelligence testing. The way we measure these constructs has evolved over time. The most significant recent change was made when we introduced the new SAT. A number of us think that we need to evolve these constructs further. One goal might be to find ways to measure these constructs so that test scores show less impact on subgroups of our testing population.

One way we try to ensure the continued relevance of the test content is by the involvement of committees and councils.

SAT Test Development Process

- SAT Advisory Boards and Committees
  - SAT Committee
  - Advisory Panel on Student Concerns
  - Council on Admissions and Guidance

The College Board has a number of Advisory Boards and Test Development Committees that impact the content of the tests they sponsor. Three committees provide input to the SAT I tests. 1) the SAT committee which is made up of high school and college faculty members and one or two members of the measurement community, 2) The Advisory Panel on Student Concerns (a committee of high school and college students) and 3) The council on Admissions and Guidance.

The SAT committee actually reviews the content of the SAT and approves changes to the test. The other two committees are mostly concerned with the operations of the test and the delivery of test related services.

The SAT II subject test committees operate in a very different way.

SAT Test Development Process

- Academic Advisory Council
- Test Development Committees

The Content of the SAT II subject tests is continually evolving to keep pace with changes with the subject matter that the tests measure. Major changes in the content of these tests is determined by the College Board Academic Advisory Council, this council typically consists of typically high school or college faculty members who are leaders in the field of teaching the particular discipline. For example, past presidents of NCTM—the national council of teachers in mathematics have been active on this advisory council.

The SAT II Test Development Committees play a very different role for these tests than the SAT committee plays for SAT I—the SAT II committees actually write and review items for the subject matter tests. The committees are made up of high school and college faculty members who are experts in their respective subject matters.

What I'd like to do next is tell you a little about the actual item writing and test assembly process.

SAT Test Development Process

- Seven Forms Developed Each Year
  - Requires 686 verbal questions and 520 math questions
- Annual Item Writing and Pretesting Requirements
  - 1,500-1,900 verbal and 1,400-1,700 math items must be written to pretest 1,300 verbal and 1,500 math questions
  - 400-500 items for verbal and math are written each year by outside item writers

ETS develops seven operational SAT I forms each year. This requires a total of 686 verbal questions and 520 math questions. About 56 verbal pretests, and 46 math pretests are developed each year—we pretest about 1300 verbal items and 1500 math questions each year—in order to pretest this number of items, we write about 1500 to1900 verbal and about 1400 to 1700 math questions each year. About 400-500 verbal and 400-500 math questions are written each year by outside

item writers. The remaining questions are written by ETS test development staff who are basically experienced teachers or subject matter experts.

SAT Test Development Process

- Pretests Administered in 30-Minute Section of Operational SAT
- Statistical Data Analyzed
  - DIF
  - Difficulty
  - Discrimination
- 15-20% Questions Lost in Pretesting

Pretests are administered along with operational forms of the SAT in a separate 30 minute section of the test. Statistical data are collected and analyzed The types of analysis that are carried out on pretests are Differential Item Functioning Analysis and Item Analysis that evaluates difficulty and discrimination indices.

Approximately 15-20% of SAT questions are lost in the pretesting process.

SAT Test Development Process

- Content Specifications
  - Item type
  - Subject matter
  - Reasoning skills
  - Gender balance
  - Minority representation (verbal only)
- Statistical Specifications
  - Difficulty distribution
  - Average r-biserial
  - DIF--no C DIF--average DIF by subgroup

Detailed specifications exist for the assembly of SAT I tests
Content Specifications
read from view graph
Statistical Specifications
read from view graph

SAT Test Development Process

- Two Years to Assemble New Form of SAT I
- Over 200 Steps in Current Test- Development Process
- Very Labor-intensive and Expensive
- Goal of Test Creation Reengineering Process
  - 40% Reduction in cost for item acquisition and assembly

It takes approximately two years to assemble a new SAT. There are over 200 steps in the current test development process that include, writing, reviewing, and pretesting items, and assembling, reviewing and printing tests.

You can see that this process is very long and labor intensive and consequently very expensive. As I mentioned earlier, we are currently reengneering our test creation process with the goal of a 40% reduction in our costs of item acquisition and test assembly.

I'd like to say a few words about the new test creation process next.

The Problem

- ETS Needs to Remain Competitive in a Changing External Environment
- Test Creation Has Become Too Expensive, Too Slow, and Too Inflexible
- New Computer-based Tests Create Demands That the Current Process Can't Meet
- Incremental Improvement Not Enough

The problem that ETS is currently facing is that in order to remain more competitive and to help our clients remain more competitive, we have to find find ways to reduce costs, and shorten our cycle time. In addition, we need to become more flexible so that we can be more responsive to the individual needs of our customers. The need to change has been brought about by several factors—certainly competition in all segments of business has been growing, but also, we have a unique problem in that it requires about six times as many items to maintain a CBT than it does to maintain a paper and pencil testing program. The volume of items required by CBT is so great, it became clear to us early on that our current test creation process simply could not support CBT.

It also became clear to us that we needed to do more than process improvement—we needed to dramatically change the way we are creating tests.

The Solution

- Invent and Implement a *New* Process That Can
  - Reduce cost and time required to create assessments
  - Be more responsive to client/customer needs
  - Meet computer-based testing requirements
  - Improve or maintain product quality
  - Create good environment for innovation

Because we needed a dramatic and revolutionary solution to the problem we chose reengineering as our methodology for change. Our goal was to invent and implement a new process. The key issue here is that we are trying to reinvent our test creation process so it better meets our customers needs.

The next slide shows you our specific objectives for the project.

Objectives of Test Creation Reengineering

- Reduce the Overall Cost of the Process
- Create the Capability of Producing a Test Form or Pool in 4-6 Months
- Improve or Maintain Quality
- Allow Programs and Clients to make Trade-offs Between Cost, Time, and Enhanced Features Based on Competitive Needs

I should say a word about one of my goals and that is to find ways to support CBT that make less demands on item writers—it seems to me like it is a no win situation unless we can cut down on the item production needed to support these tests.

The key ways that we thought we could accomplish our goals for reengineering the test creation process are shown on the next slide.

How Will We Accomplish Our Goals?

- Better Up-front Planning for New and Ongoing Tests by Cross-functional Teams
- Create and Lock Items Before Test Assembly
- Test Assembly and Test Book Production More Automated
- Seamless Software Interface with Other Systems
- Continual Monitoring and Improvement

1. Cross functional representation of the various areas at ETS in up-front planning is critical—here are some examples of problems we've run into—we have a brand new CBT test ready to go, but the tutorial was not part of the planning process and had to be added at the last minute.

2. Locking items at the beginning of the process is at the heart of many of the cost savings—currently—multiple reviews—emphasis on test reviews—new process—emphasis on building strong pools—tests are reviewed for balance, etc, but individual items are not reviewed.

3. Tests are assembled by computer—minimal review is required.

4. Built an entire new system of test creation software to support the new process—some of it is ETS proprietary—but a lot of it uses off the shelf components so that we can take advantage of product upgrades and keep the system developing dynamically.

5. We've created an extensive system of metrics to go with our new process to help us continually monitor the effectiveness of the system and plan improvements.

We are currently near the end of our three year reengineering project, just about in the middle of what we are referring to as Phase III.

Where Are We Now?

- Reengineering a Three-Phase Project
- First Phase
- Design and Development
- Second Phase
- Software and Process Development
- Third Phase
- Implementation

The reengineering project began in 1996 with an extensive design phase. The second phase was very heavy on the development of the process and supporting software. Phase III is the implementation phase. This phase should conclude next December and at that point the new process should be completely rolled out. Basically, what we are up to our ears in now is training staff in the new process and motivating them to make the extensive changes that are required by this process.

Scoring of the SweSAT

The scoring of the restructured SweSAT was discussed. Results of SAT as well as results of PET are given by two scores, one verbal and one mathematical, while the results of SweSAT are given in one combined normed score only, which is based on the raw score i.e. the total number of correct answers.

Testtakers have complained that the subtest WORD is given too much importance. The subtest contains 40 items which is almost one third of the total number of items, but the time for the subtest is only 15 minutes out of the 4 hours and 10 minutes total testtime. Since each cor-

rect answer is given one point the WORD subtest can give 40 out of 122 possible points.

The decision about weighting of subtest scores is actually a political matter. Otherwise the decision should be theoretically based. When validity results are lacking one way to go would be to ask university professors what abilities they regard as most important for study success. Another way to go would be to control for the standard deviations of the subtests before they are combined

A first factor explaining only 10.2 per cent of the variance is rather low – about 20 per cent is more common. This result indicates that SweSAT is a multidimensional test.

The reliability of the subtest READ is comparatively low, but in the absence of proper criterion data, the reliability should not be decisive for the weighting.

More factoranalytical studies of the test are necessary.

Differences between subgroups on SweSAT 1996B

Kristian Ramstedt

In this presentation I am going to demonstrate differences between some subgroups on the different subtests in the Swedish scholastic aptitude test (SweSAT) and on the total test.

In 1996 the composition of SweSAT was changed. One subtest was excluded (the general information test) and the number of items in some of the other tests were changed (see table 1 were “MAX” gives the number of items in the different subtests).

Table 1 *Differences in raw scores, standard deviations and effect sizes, number of students (10% random sample) and number of items for tests 1995B and 1996B*

Year	Score means			Std. Dev.		F-M	Number			
test	Fem	Male	F-M	Fem	Male	pool	Fem	Male	F-M	MAX
95B										
GI	18,60	19,85	-1,25	3,92	3,91	**-0,32**	2845	2447	398	**30**
DTM	11,09	13,38	-2,29	3,50	3,36	***-0,67***	2845	2447	398	**20**
ERC	15,72	17,05	-1,33	5,02	4,85	***-0,27***	2845	2447	398	**24**
READ	14,34	14,99	-0,65	3,88	3,84	***-0,17***	2845	2447	398	**24**
DS	9,90	12,39	-2,49	3,67	3,81	***-0,67***	2845	2447	398	**20**
WORI	19,66	20,34	-0,68	5,51	5,35	***-0,13***	2845	2447	398	**30**
TOT	89,30	98,00	-8,70	19,74	19,28	***-0,45***	2845	2447	398	**148**
96B										
DTM	10,87	13,28	-2,41	3,56	3,37	***-0,69***	3169	2442	727	**20**
ERC	12,36	13,68	-1,32	4,28	4,04	***-0,32***	3169	2442	727	**20**
READ	13,10	13,47	-0,37	3,29	3,17	***-0,11***	3169	2442	727	**20**
DS	12,42	14,98	-2,56	3,98	3,92	***-0,65***	3169	2442	727	**22**
WORI	23,61	24,04	-0,43	6,89	6,80	***-0,06***	3169	2442	727	**40**
TOT	72,36	79,45	-7,09	17,03	16,08	***-0,43***	3169	2442	727	**122**

Table 1 shows the differences between females and males on the 95B and 96B tests. The column marked “F-M pool” shows the effect size in pooled standard deviations.

As can be seen from table 1 males are doing better on all subtests. The change of format did not affect the differences significantly. The overall difference is – 0.45 in 95B and – 0.43 in 96B. In both tests the subtests DTM (diagrams, maps and tables) and DS (data sufficiency) show the largest gender differences. The change of format, however, did not have any significant impact on the gender differences.

I will also show some diagrams (box plots) in order to show differences between different age groups as well as gender differences and also some box plots showing differences between test takers of the same age but from the different programs in upper secondary school.

Different age groups

When SweSAT was introduced in the late seventies it was aimed for people with work experience but without the proper educational background, the so called 25:4 group, meaning that the test taker had to be at least 25 years old and to have 4 years of work experience. In 1992 the rules were changed and from then on there are no limitations for

participating in the SweSAT. This change of the rules increased the number of participants more than tenfold. However, there are still a lot of older participants and it might be interesting to compare results for different age groups on different subtests and also gender differences in different age groups.

Figure 1 shows the results on the 1996 B DTM subtest. Males are outperforming females in all age groups. It can also be seen that the results are decreasing with higher ages.

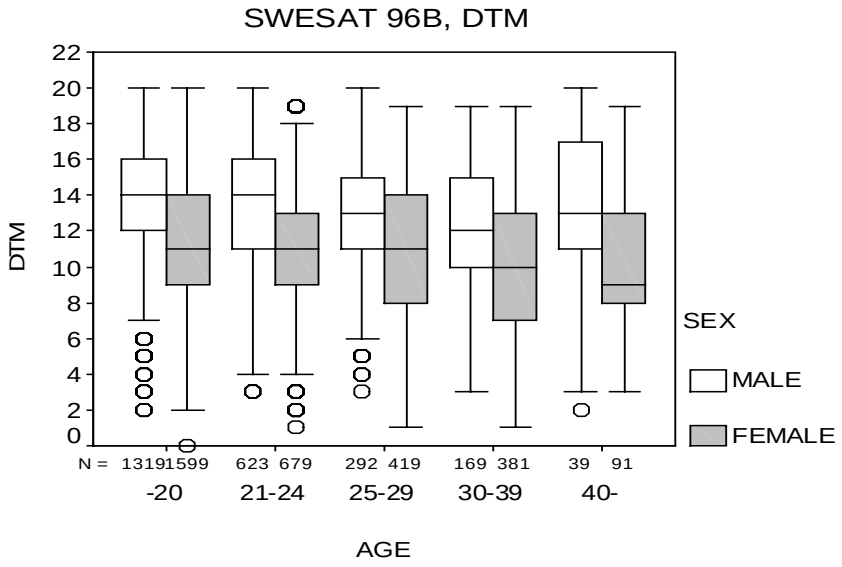


Figure 1 *The 1996B DTM test results for different age groups and for females and males.*

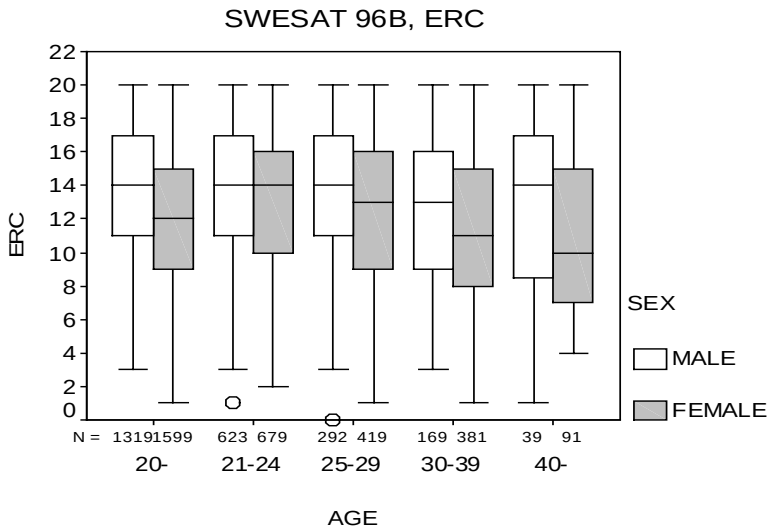


Figure 2 *The 1996B ERC test results for different age groups and for females and males.*

Figure 2 shows the same pattern as Figure 1 even if the gender differences are somewhat smaller for the ETC subtest. Figure 3 shows the result for the READ subtest.

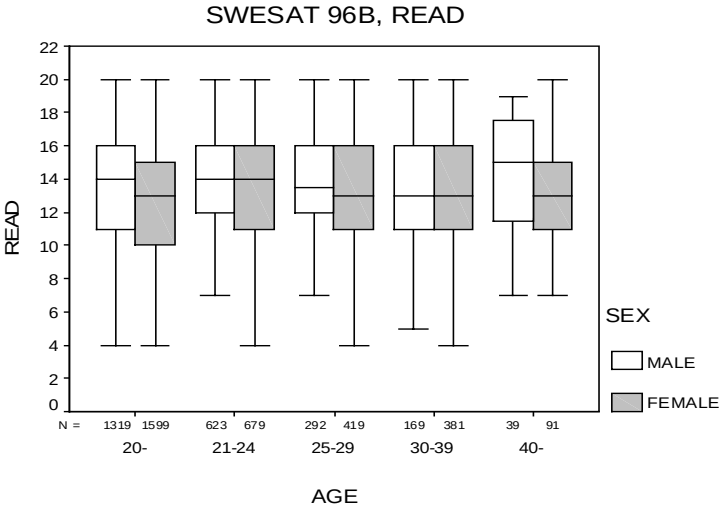


Figure 3 *The 1996B READ test results for different age groups and for females and males.*

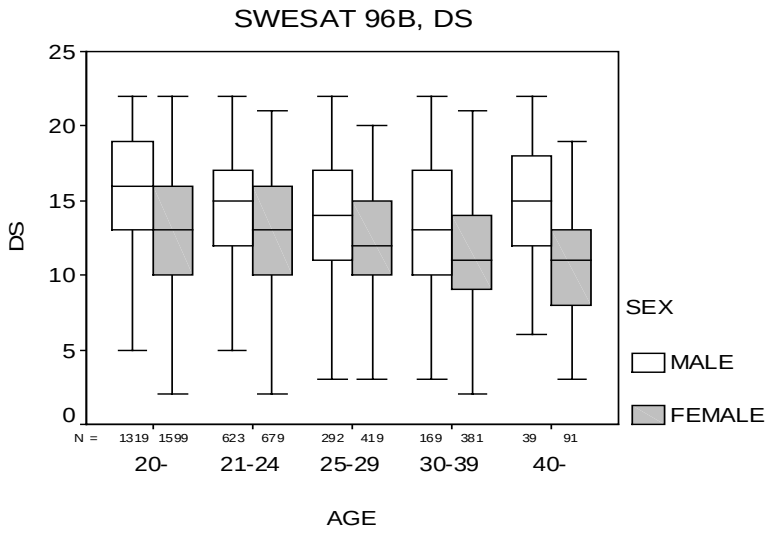


Figure 4 *The 1996B DS test results for different age groups and for females and males.*

The READ subtest shows small or no gender differences. The age differences are also much smaller than for the two other subtests.

The DS subtest shows a pattern very similar to the DTM subtest. Large gender differences and a decreasing result for older age groups.

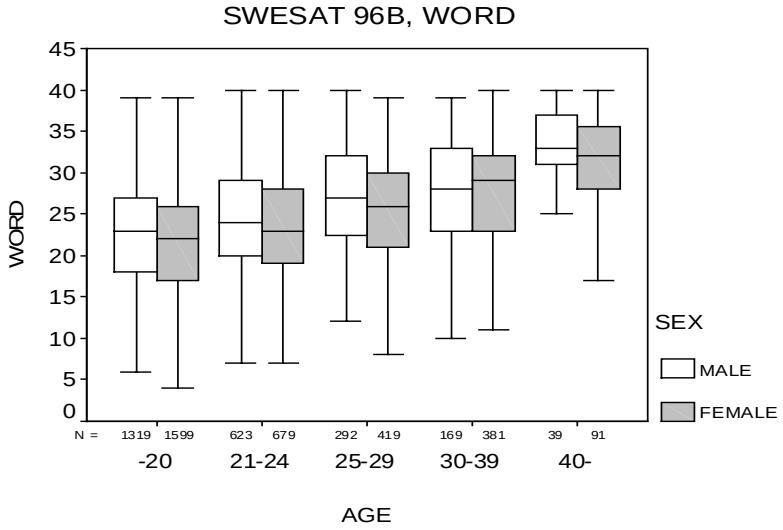


Figure 5 *The 1996B WORD test results for different age groups and for females and males.*

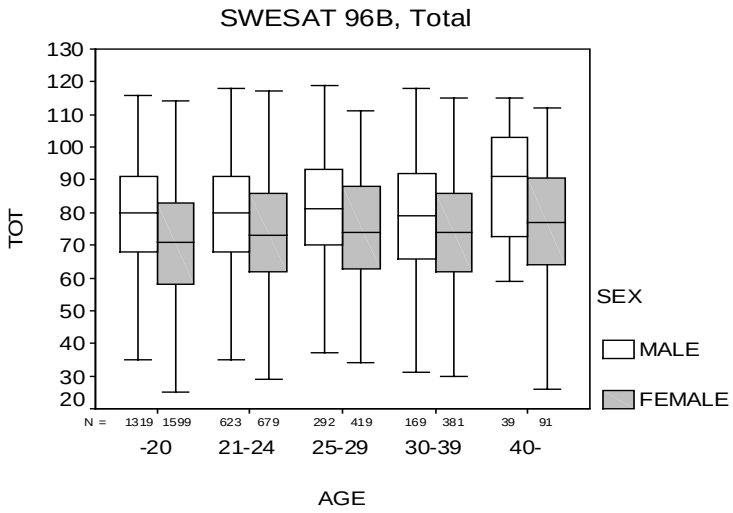


Figure 6 *The 1996B TOTAL test results for different age groups and for females and males.*

The WORD subtest on the other hand shows an increasing result for older age groups and quite small gender differences but still a small advantage for males.

When all the subtest results are added the picture looks like Figure 6. There are still significant gender differences in all age groups but the differences between different age groups are very small. Whether or not this is fair is of course a matter of discussion but at least there are no obvious differences like those for gender.

One possible cause of the gender differences might be different educational backgrounds for females and males. In order to condition the results on educational background we will now look at the results of female and male test takers born in 1978 and studying on the same programs in the upper secondary school. It will then be possible to compare results for females and males from the same and from different programs with each other.

Groups with different educational background.

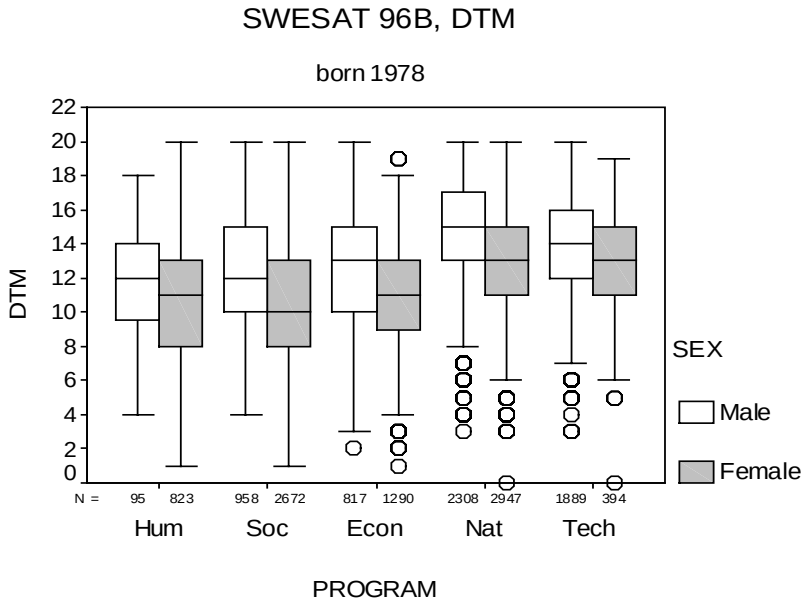


Figure 7 *The 1996B DTM test results for different program groups and for females and males.*

All test takers in this comparison are born in 1978. They are studying on one of the five theoretical programs in upper secondary school, humanistic (Hum), social science (Soc), economics (Econ), natural science (Nat) and technical engineering (Tech). Figure 7 shows the results on the DTM subtest.

As can be expected the natural science and technical engineering students are doing best on the DTM subtest. We can also note that males are doing significantly better within all programs.

The English reading comprehension subtest the picture is a little bit more mixed. Females from the humanistic program are, not surprisingly, doing well on this test. The same goes for the Nat - students. More surprising is maybe that the males from the social science program are doing so well while females are not. The gender differences in favor of males are present in all programs but the humanistic.

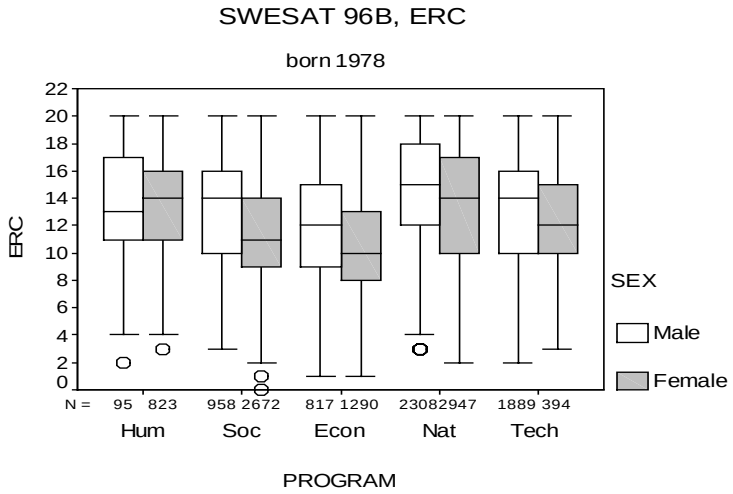


Figure 8 *The 1996B ERC test results for different program groups and for females and males.*

The results on the Swedish reading comprehensive test READ are shown in Figure 9.

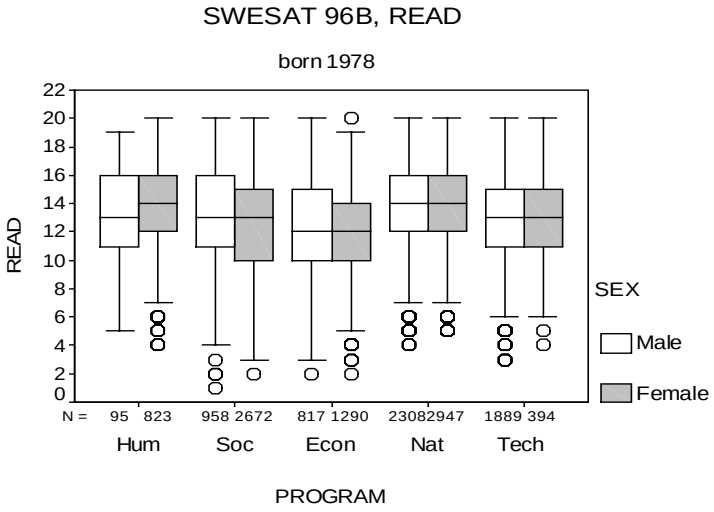


Figure 9 *The 1996B READ test results for different program groups and for females and males.*

In the READ test the picture looks much nicer than in the former subtests. The gender differences within the programs are very small. Since box plots are showing medians in integers, however, there might be differences in the means as Figure 10 shows.

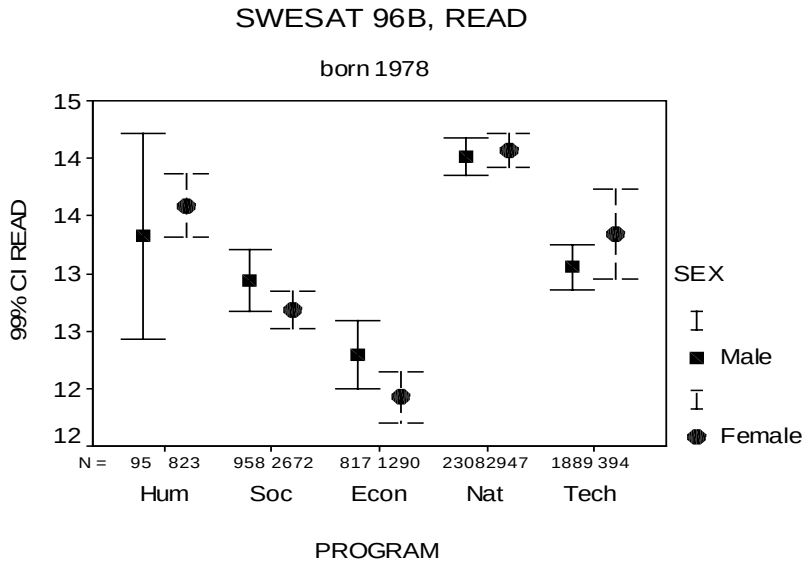


Figure 10 *The 1996B READ test results for different program groups and for females and males. Means and 99% confidence intervals.*

As we have seen earlier DS is a subtest with rather large gender differences. Figure 11 shows this is the case here too. The picture is very much the same as in Figure 7 showing the DTM results.

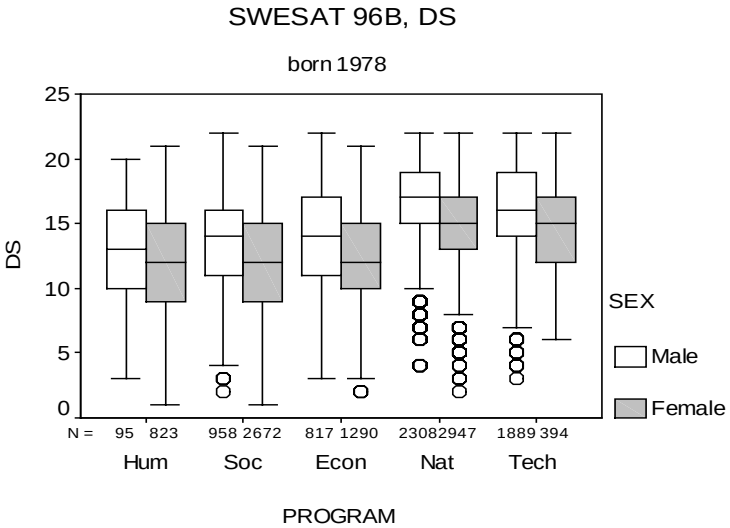


Figure 11 *The 1996B DS test results for different program groups and for females and males.*

The Word subtest is shown i Figure 12.

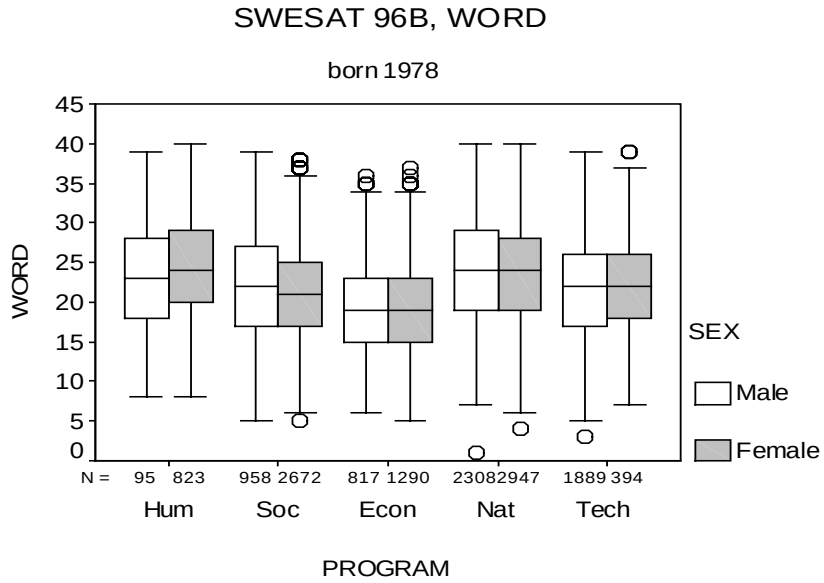


Figure 12 *The 1996B WORD test results for different program groups and for females and males.*

Here the gender differences are negligible within the programs and the program differences are what could be expected.

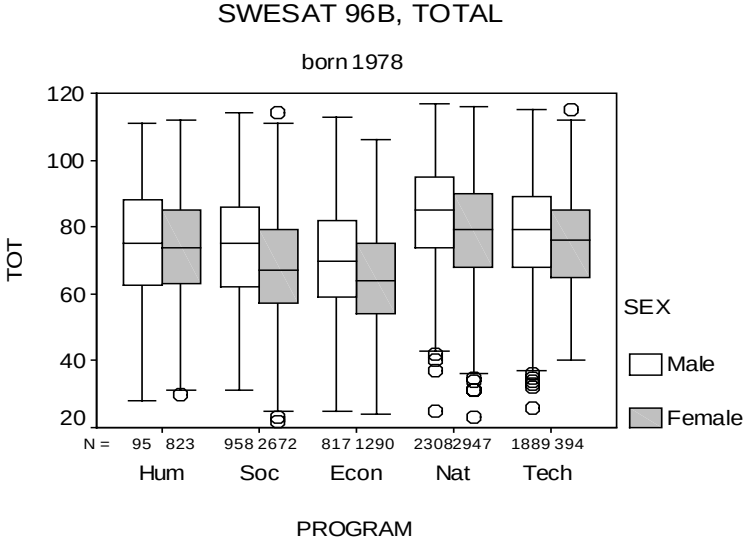


Figure 13 *The 1996B total test results for different program groups and for females and males.*

As can be seen from Figure 13 there are smaller or larger gender differences in favor of the males in all programs. The natural science and technical engineering students are doing best, mainly because of their advantage on the quantitative subtests DTM and DS.

Even if the educational backgrounds are the same we still do not have any certain external criteria to validate the SweSAT results against. If we knew the grades for the students we could use them. What we do know (internal material) is that the boys are outperforming the girls on the national test in physics, chemistry and mathematics for the natural science and technical engineering programs (most significantly in physics and less in mathematics) for the age group studied here. We also know that the difference between boys and girls in test result and grades (measured in effect sizes) are not the same. There is a difference of about 0.10 pooled standard deviations in favor of girls if you look at grades, in favor of boys if you consider test results more reliable.

In more verbal subjects, however, girls are getting higher grades and you could think that their higher verbal ability should compensate for their losses on the quantitative subtests. But since the participants in the SweSAT are selfselected, as are the students in the different programs, it is very difficult to draw any definite conclusions from a material like the one presented here.

Item analysis based on item response theory and on classical test theory. A comparison

Christina Stage

Ever since SweSAT was first taken into use in 1997, the development and assembly of the test, as well as equating of forms from one administration to the next, has been based on classical test theory (CTT).

The statistics which are used in the item analysis are:
p-values of the items
p-values of the distractors
p-values of males and females
biserial correlations (r_{bis})
(the item test regression)

Since spring 1996, pretesting of new items for SweSAT has been performed in connection with the regular test administration, which

means that the examinee sample on which pretesting is performed is a sample from the true examinee population and it contains 1500 examinees as a minimum. This new procedure for pretesting makes it possible to use IRT for item analysis and compilation of new test versions.

In the SweSAT given in spring 1997, the subtest WORD contained 20 items, which had been pretested on five different samples from the examinee population in spring 1996; the subtest ERC contained 14 items which had been pretested on five different samples, and the subtest READ contained 16 items which had been pretested on eight different samples in spring 1996.

For these common items a comparison was made between item statistics from the CTT framework and item parameters estimated within the IRT framework. Specifically the following questions were addressed:

1. How accurate are the predictions made from pretest data to regular test data within the IRT framework compared to the same predictions made within the CTT framework?
2. How do item difficulty indices from CTT compare with item difficulty parameters estimated by IRT?
3. How do item discrimination indices from CTT compare with item discrimination parameters estimated by IRT?

Method

Classical test theory

For the items which had been pretested in spring 1996 and were used in the regular test in spring 1997, the p-values and the biserial correlations (r_{bis}) were calculated. The same indices were calculated for the corresponding items in the pretest data and the values were compared.

Item Response Theory

The five WORD pretest combinations, the five ERC pretest combinations and the eight READ pretest combinations in spring 1996 were run in BILOGW together with the regular WORD, ERC and READ subtests from spring 1996, and the a-, b- and c-parameters were estimated. The WORD subtest, the ERC subtest and the READ subtest

from spring 1997 were run (separately) in BILOGW and the three item parameters were estimated. The parameter estimates for the corresponding items were noted and compared.The ICCs for the corresponding items were also compared.

Results

The WORD subtest

Classical test theory

For the 20 common WORD-items the correlation between p-values from pretest and regular test was was r =.93 and the Spearman rank correlation (ρ) = .92; for the same p-values transformed to delta the correlation was r =.93. In Figure 1 (left) the p-values from the regular test spring 1997 are plotted against the p-values from the pretest.

The correlation between r_{bis} was r = .81 and ρ = . 72. A plot of r_{bis} is shown in Figure 1 (right).

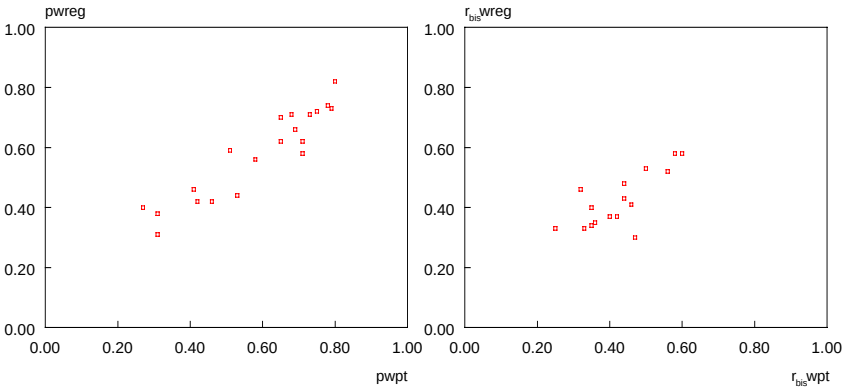


Figure 1 *Plots of p-values (left) and r_{bis} (right) from the regular WORD subtest against p-values and r_{bis} from the pretest versions.*

Item Response Theory

The correlation between the b-values estimated on pretest and regular test data was r = .92 and ρ = .92. A plot of the b-values is shown in Figure 2 (left).

The correlation between a-values estimated on pretest and regular test data was r = .74 and ρ = .60. The plot is shown in Figure 2 (right).

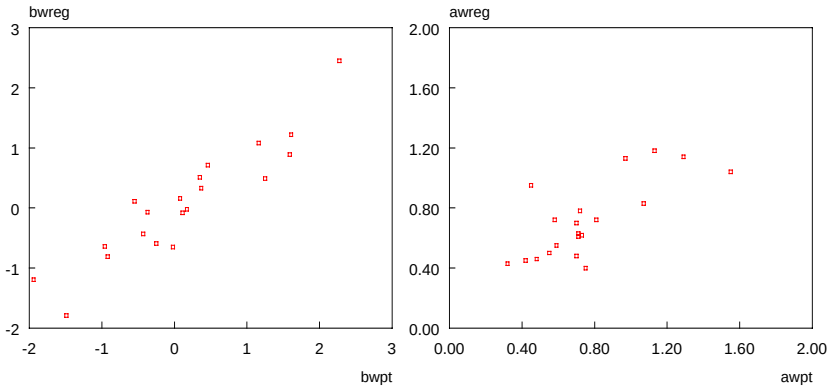


Figure 2 *Plot of b-values (left) and a-values (right) from the regular test against b- and a-values from the pretest.*

The correlation between c-values was r = .74

As for model data fit of the IRT model used (three parameter logistic model) none of the 20 items was identified as misfitting at α = .01

Item Characteristic Curves of some WORD-items

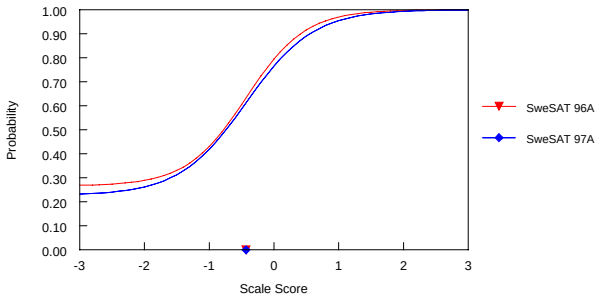


Figure 3 *ICCs for item No 1 in the Spring 1997 regular WORD subtest and item No 8 in the pretest in Spring 1996.*

In this item two of the distractors had been changed after the pretest. As may be seen in Figure 3 the two ICCs correspond very well, the b-values were exactly the same (-.43), while the a-value in the pretest was 1.29 and in the regular test 1.14. The p-value for this item in the pretest was p = .73 and in the regular test p = .71; the r_{bis} in the pretest was .60 and in the regular test it was .58.

On the whole the results from CTT and IRT correspond very well and according to both analyses this item seems to work in the same way in the pretest as in the regular test.

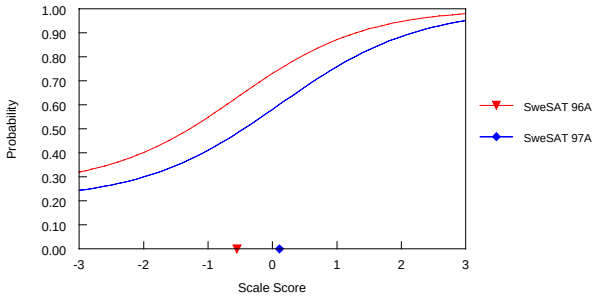


Figure 4 *ICCs of item No 15 in the regular test and No 36 and the pretest.*

In item No 15 one distractor had been changed between pretest and regular test. The b-value had increased from -.55 to .11, while the a-value had decreased to a very small extent (from .59 to .56). The p-value had decreased from .71 to .52 and the r_{bis} from .40 to .37.

Again the conclusions are the same, the item was more difficult and somewhat less discriminating in the regular test than in the pretest.

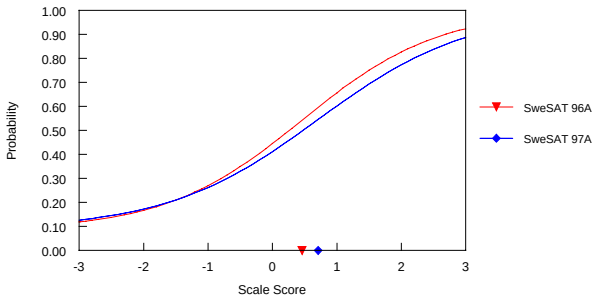


Figure 5 *ICCs of item No 19. In the regular test and No 5 in the pretest.*

In item No 19 nothing had been changed, but the b-value had increased from .46 to .71, while the a-value had decreased from .55 to .50.The p-value had decreased from .46 to .42 and the r_{bis} from .42 to .37.

According to both analyses this item was a bit more difficult and less discriminating in the regular test than in the pretest.

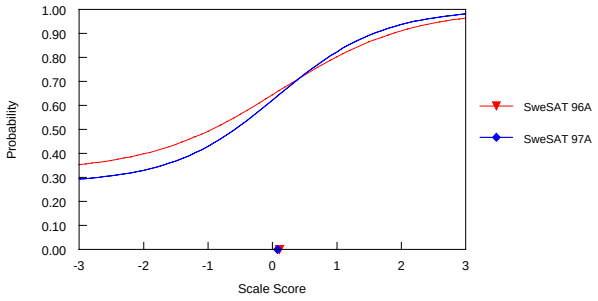


Figure 6 *ICCs of item No 23 in the regular test and No 16 in the pretest.*

In item No 23 no changes had been made and the b-value was almost the same (.11/.08) in the regular test as in the pretest but the a-value had increased from .58 to .72.The p-value had decreased from .65 to .62 and the r_{bis} had increased from .35 to .40.

According to IRT this item was imperceptibly easier but had better discrimination power in the regular test than in the pretest. According to CTT the item was a bit more difficult but also better discriminating in the regular test than in the pretest.

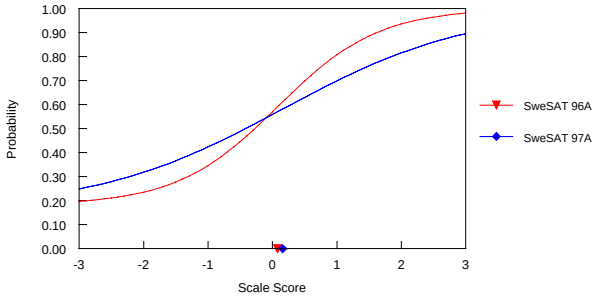


Figure 7 *ICCs of item No 24 in the regular test and No 38 in the pretest.*

In item No 24 no changes had been made but the b-value had increased slightly from .08 to .16, while the a-value had decreased from .75 to .40 from pretest to regular test. The p-value had decreased from .58 to .56 and the r_{bis} had decreased from .47 to .30.

According to both analyses the item had become a little more difficult but less discriminating in the regular test than in the pretest.

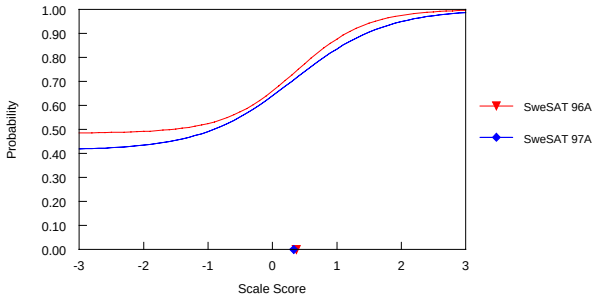


Figure 8 *ICCs of item No 27 in the regular test and No 24 in the pretest.*

In item No 27 no changes had been made but still the b-value had decreased from .37 to .33 and the a-value had decreased from 1.07 to .83 from pretest to regular test. The p-value had decreased from .69 to .66 and the r_{bis} had decreased slightly (from .36 to .35) from pretest to regular test.

Hence according to IRT this item was somewhat easier in the regular test than in the pretest, while according to CTT the item was somewhat more difficult in the regular test. According to both analyses the discrimination power had decreased to a small extent from pretest to regular test.

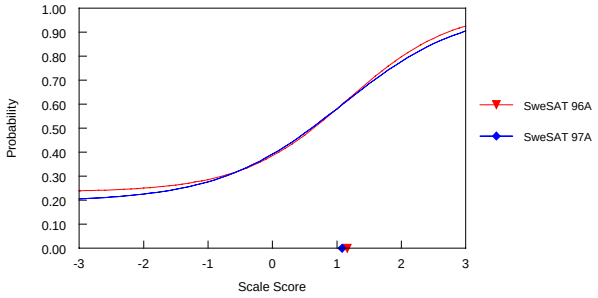


Figure 9 *ICCs of item No 29 in the regular test and No 4 in the pretest.*

In item No 29 no changes had been made but the b-value had decreased slightly from 1.16 to 1.08 and the a-value had decreased from .71 to .61 from pretest to regular test. The p-value was the same (.42) in the pretest as in the regular test and so was the r_{bis} (.33).

Hence according to IRT the item was somewhat easier and poorer discriminating in the regular test than in the pretest. According to CTT

the difficulty level was exactly the same and so was the discrimination power.

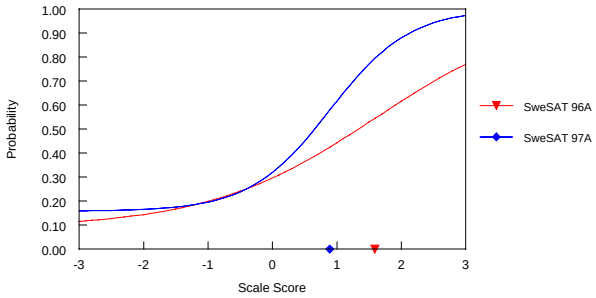


Figure 10 *ICCs of item No 35 in the regular test and No 5 in the pretest.*

In item No 35 one distractor had been changed and the the b-value had decreased substantially (from 1.59 to .89) but the a-value had increased (from .45 to .95) between pretest and regular test. The p-value had increased from .31 to .38 and the r_{bis} had increased from .32 to .46 from pretest to regular test.

Hence according to both analyses this item was easier but more dicriminating in the regular test than in the pretest.

Comparison between CTT and IRT

The correlation between estimated b-values and p-values was r = -.93 and ρ = -.95 for the pretest items and r = -.94 and ρ = -.96 for the regular test items. The difficulty indices for the regular test items are plotted in Figure 11 (left).

The correlation between estimated a-values and r_{bis} was r = .62 and ρ = .65 for the pretest items and r = -.65 and ρ = .64 for the regular test items. In Figure 11 (right) the plot of a-values against r_{bis} for the regular test items is shown.

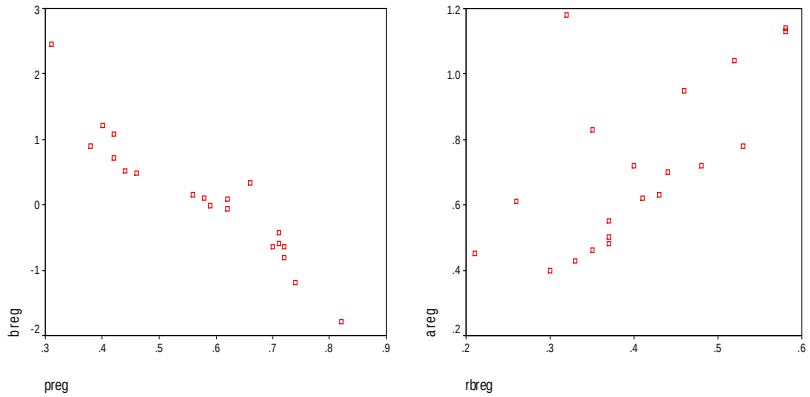


Figure 11 *Estimated b-values plotted against p-value (left) and estimated a-values plotted against r_{bis} (right) for the regular WORD test items.*

The ERC subtest

Classical test theory

The correlation between p-values of the items in the pretest versions and p-values of the corresponding items in the regular test version of ERC was r = .86 and ρ = .87. A plot of the p-values is shown in Figure 15 (left).

The correlation between r_{bis} of the items in the pretest versions and the corresponding items in the regular test version was r = .57 and ρ = .64; a plot of the r_{bis} is shown in Figure 15 (right).

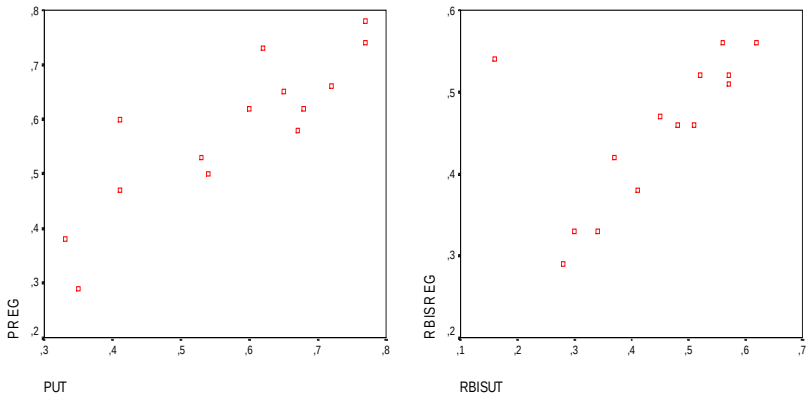


Figure 12 *Plot of p-values(left) and r_{bis} (right) for regular test items against p-values and r_{bis} for pretest items.*

The items deviating most from the linear regression lines were No 5 and No 9, which were also the two items in the ERC subtest that had been changed between pretest and regular test. When items No 5 and No 9 were removed, the correlation between p-values for the remaining 12 items was r = .95 and ρ = .94 and the correlation between r_{bis} was r = .96 and ρ = .94.

Item Response Theory

The correlation between b-values estimated on regular test data and pretest data was r = .88 and ρ = .83 A plot of the b-values is shown in Figure 13 (left).

The correlation between a-values estimated on regular test data and a-values estimated on pretest data was r = .34 and ρ = .58. The plot of a-values is shown in Figure 13 (right).

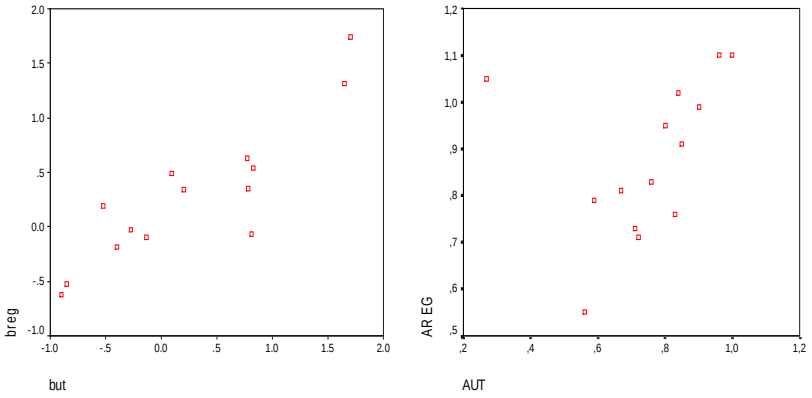


Figure 13 *Plot of b-values (left) and a-values (right) estimated on the regular test items against values estimated on the pretest items.*

The most deviating items were No 5 and No 9. When these items were removed the correlation between b-values increased to r = .96 and ρ = .94 and the correlation between a-values to r = .82 and ρ = .80.

The correlation between c-values estimated on pretest data and c-values estimated on regular test data was r = .80 and ρ = .58.

The assessment of model data fit showed that for two items, No 9 and No 14 there was a model data misfit which was significant at α = .01 level.

Item characteristic curves of some ERC items.

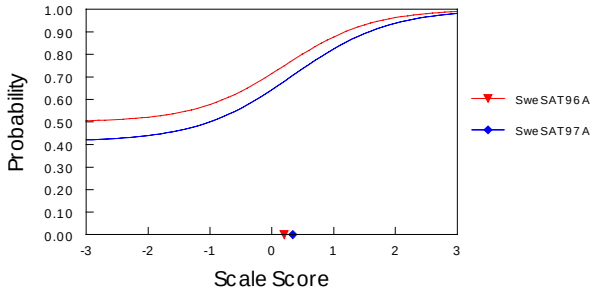


Figure 14 *ICCs of item No 2 in the regular test and No 2 in the pretest.*

For item No 2 the estimated b-value had increased from .20 to .34 while the estimated a-value had decreased from .83 to .76. For the same item the p-value had decreased from .72 to .66 and the r_{bis} from .34 to .33.

According to both analyses the item was slightly more difficult and less discriminating in the regular test than in the pretest.

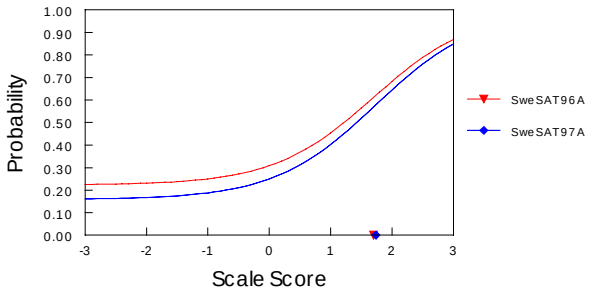


Figure 15 *ICCs of item No 3 in the regular test and No 3 in the pretest.*

For item No 3 the estimated b-value had increased from 1.70 to 1.74 and the estimated a-value had decreased from .72 to .71. For the same item the p-value had decreased from .35 to .29 and the r_{bis} had changed from .28 to .29.

According to both analyses the item was imperceptibly more difficult but had about the same discrimination power in the regular test as in the pretest.

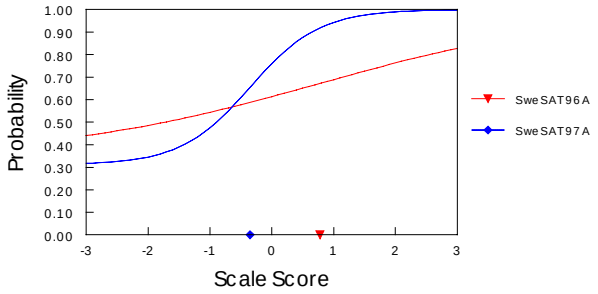


Figure 16 *ICCs of item No 5 in the regular test and No 5 in the pretest.*

Item No 5 was one of the two items in the ERC subtest which had been changed between pretest and regular test. The b-value had decreased from .78 to .35, while the a-value had increased from .27 to 1.05. For the same item the p-value had increased from .62 to .73 and the r_{bis} from .16 to .54.

According to both analyses the item was easier and better discriminating in the regular test than in the pretest. The change seems to have improved the item.

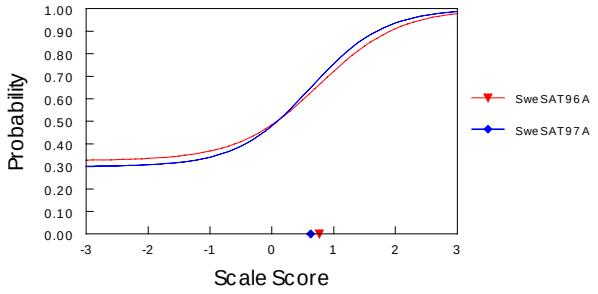


Figure 17 *ICCs of item No 8 in the regular test and No 3 in the pretest.*

For item No 8 the b-value had decreased from .77 to .63 while the a-value had increased from .90 to .99. For the same item the p-value was the same (.53) in the pretest as in the regular test while the r_{bis} had increased from .37 to .42.

According to IRT this item was slightly easier in the regular test than in the pretest while according to CTT the difficulty level was the same. According to both analyses the discrimination was somewhat better in the regular test than in the pretest.

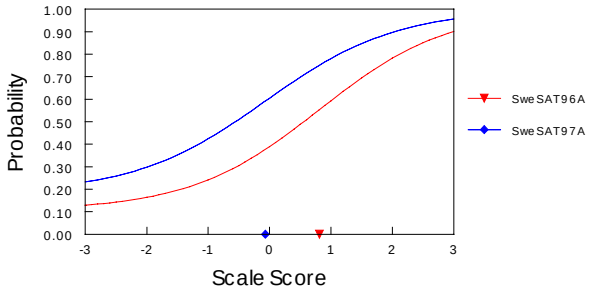


Figure 18 *ICCs of item No 9 in the regular test and No 11 in the pretest.*

This was the other item in the ERC subtest which had been changed between pretest and regular test. For this item the b-value had decreased from .81 to -.07 while the a-value was about the same (.56 and .55 respectively). For the same item the p-value had increased from .41 to .60 while the r_{bis} had decreased from .41 to .38 from pretest to regular test.

Hence according to both analyses the item was easier and slightly less discriminating in the regular test than in the pretest. For this item the change does not seem to have caused any improvement.

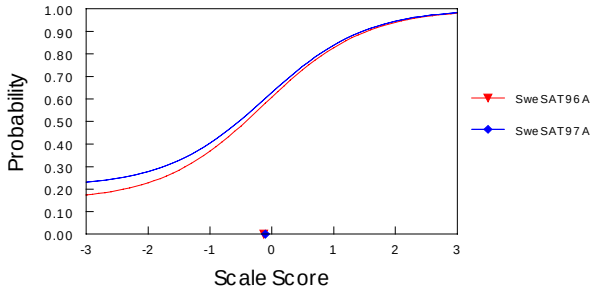


Figure 19 *ICCs of item No 12 in the regular test and No 14 in the pretest.*

For item No 12 the b-values were almost the same (-.13 and -.10 respectively) in the pretest as in the regular test and so were the a-values (.71 and .73 respectively). For the same item the p-values too were very close (.60 and .62 respectively) while the r_{bis} had decreased from .51 to .46 from pretest to regular test.

According to both analyses the difficulty level of the item was about the same in the pretest and the regular test but according to IRT the

discrimination power was slightly better in the regular test and according to CTT it was slightly poorer.

Comparison between CTT-based and IRT-based item indices

The correlation between the IRT estimated b-values and the CTT calculated p-values for the 14 common items was r = -.90 for the pretest items as well as for the regular test items (ρ = -.88 and -.82 respectively). For the regular ERC test items a plot of the difficulty statistics according to the two theories is shown in Figure 20 (left).

The correlation between r_{bis} and estimated a-values was r = .74 for pretest items and r = .76 for regular test items (ρ = .66 and .85 respectively). A plot of item discrimination indices for the regular test items is shown in Figure 20 (right).

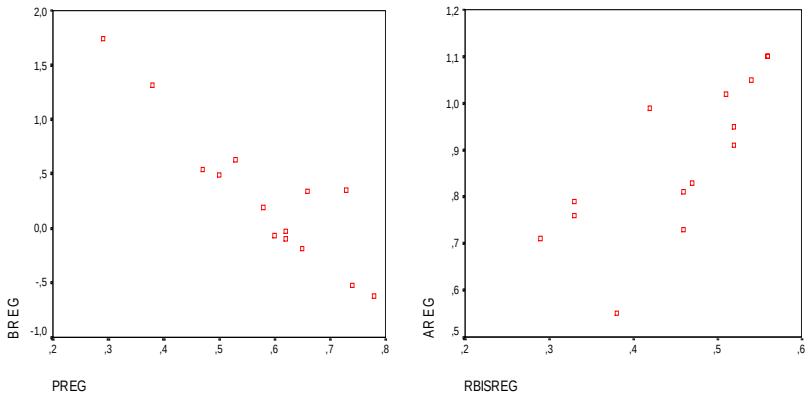


Figure 20 *Estimated b-values plotted against p-values (left) and estimated a-values plotted against r-bis (right).*

The READ subtest

Classical Test Theory

The correlation between pretest and regular test p-values for the 16 READ items was r = .78 and ρ = .82 and in Figure 21 (left) a plot of the p-values is shown.

The correlation between r_{bis} was r= .66 and ρ= .56 and a plot of the r_{bis} is shown in Figure 21 (right).

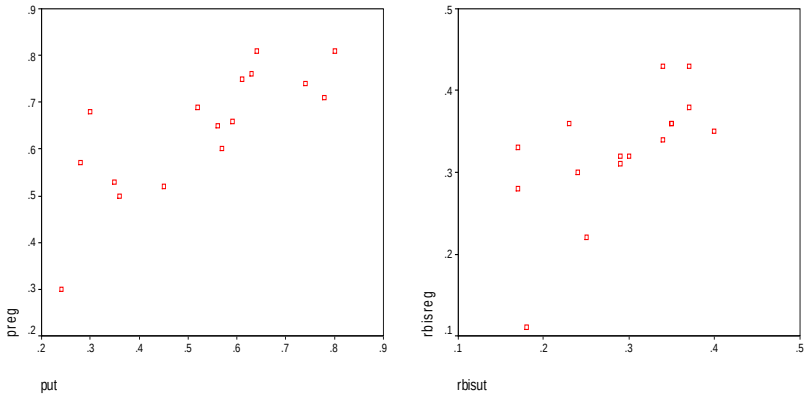


Figure 21 *Plot of p-values (left) and r_{bis} (right) from the regular READ items against p-values and r_{bis} from the pretest items.*

Item Response Theory

The correlation between b-values of items estimated on the regular test and the same items in the pretest versions was r = .55 and ρ = .56. A plot of the b-values is shown in Figure 22 (left).

The correlation between a-values was r = .54 and ρ = .51 and a plot of the a-values is shown in Figure 22 (right).

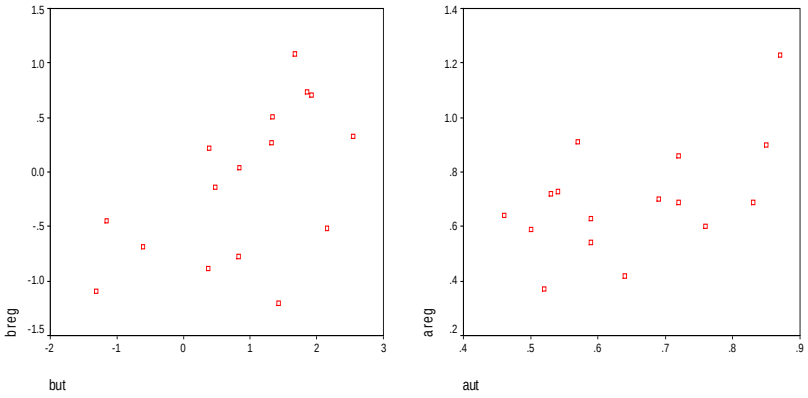


Figure 22 *Plot of b-values (left) and a-values (right) estimated on the regular READ test against b-values and a-values estimated on the pretest versions*

The correlation between c-values was r = .61 and ρ = .75.

The assessment of model data fit showed that for one item, No 11, there was a model data misfit which was significant at $\alpha = .01$ level.

Item characteristic curves of some READ items

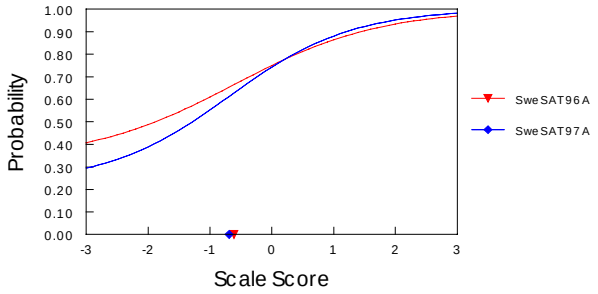


Figure 23 *ICCs of Item No 5 in the regular test and No 5 in the pre-test.*

For item No 5 the b-value had decreased slightly from -.61 to -.69 and the a-value had increased from .50 to .59. For the same item the p-value was the same (p - .74) in the regular test as in the pretest, while the r_{bis} had increased slightly from .30 to .32.

According to both analyses the item had about the same difficulty level but somewhat higher discrimination power in the regular test than in the pretest.

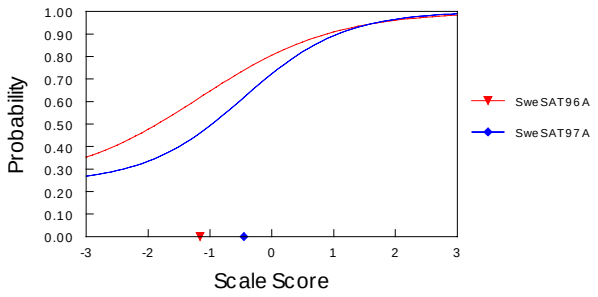


Figure 24 *ICCs of item No 7 in regular test and No 6 in the pretest.*

For this item the b-value had increased from –1.16 in the pretest to -.45 in the regular test and the a-value had increased from .54 to .73. For the same item the p-value had decreased from .78 to .71 and the r_{bis} from .37 to .38 between pretest and regular test.

According to both analyses this item was somewhat more difficult and better discriminating in the regular test than in the pretest.

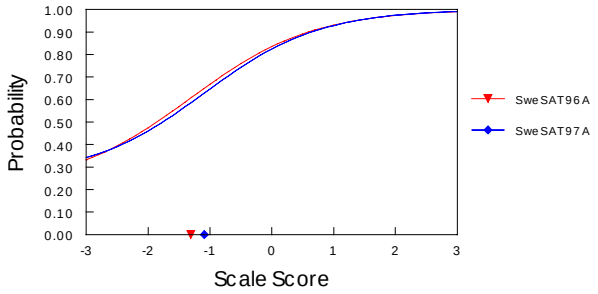


Figure 25 *ICCs of item No 8 in the regular test and No 8 in the pretest.*

For this item the b-value had increased from –1.31 to –1.09 and the a-value from .59 to .63 between pretest and regular test. For the same item the p-value had increased from .80 to .81 while the r_{bis} had decreased from .40 to .35.

Even though the differences are very small, the changes are estimated differently by CTT and IRT as the discrimination of the item according to IRT is higher and according to CTT is lower in the regular test than in the pretest.

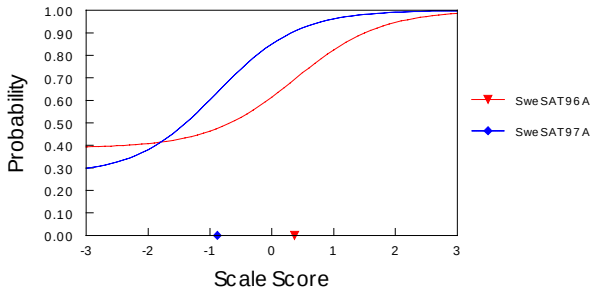


Figure 26 *ICCs of item No 9 in the regular test and No 17 in the pretest.*

For this item the b-value had decreased from .37 to -.88 and the a-value had increased from .64 to .90 from pretest to regular test.

For the same item the p-value had increased from .64 to .81 and the r_{bis} from .37 to .43.

Even though the differences between pretest and regular test are substantial they are judged in the same way by CTT and IRT; the item is easier and better discriminating in the regular test than in the pretest.

Comparison between CTT and IRT

The correlation between estimated b-values and p-values was r = -.90 and ρ = -.88 for the pretest and r = -.92 and ρ = .95 for the regular test items. A plot of the difficulty indices for the regular test is shown in Figure 27 (left).

The correlation between estimated a-values and r_{bis} was r = . 35 and ρ = .32 for the pretest items and r = .78 and ρ = .68 for the regular test items. A plot of the regular test items is shown in Figure 27 (right).

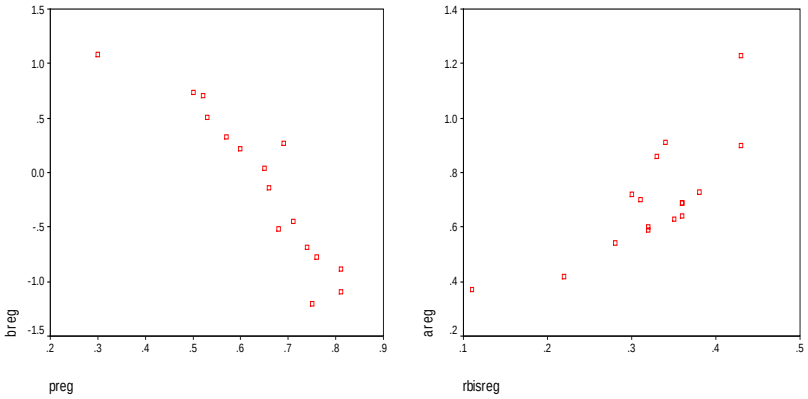


Figure 27 *Plot of estimated b-values against p-values (left) and estimated a-values against r_{bis} (right) for 16 regular READ items*

Discussion

The agreement between results from item-analyses performed within the IRT and the CTT framework is very good. For most of the items analysed the conclusions about the change between pretest and regular test regarding difficulty level as well as discrimination were the same. For only very few items the conclusion about both difficulty and discrimination differred. For all three subtests the correlations between IRT-difficulties and CTT-difficulties were higher than .90.

One problem when analysing the stability of the item parameters, for real data, is that pretesting has two purposes. One aim is to get information about the difficulty level and the discrimination power of the items in order to be able to compile parallel tests.The other purpose is to make sure that all the items function in a satisfactory way. Distrac-

tors which did not work in the pretest were changed before the item was included in the regular test, and the effect of such changes is not always possible to foresee. Anyhow, the changes mean that these items are not exactly the same in the pretest version as in the regular test. Another problem is that the order of presentation in the pretest booklets may differ from the order in the regular test. Even though the subtests are not actually speeded, changes in the order of presentation may still change the item in some way.

The overall conclusion from this study is that the prediction from pretest to regular test data is satisfactory and the major part of the discrepancy in the prediction can be explained by changes of distractors or different order of presentation of the items. This conclusion, however, is true for both analyses regardless of the theoretical framework.

What is important when compiling a test like SweSAT, is to be able to predict the difficulty level of the regular test from the pretest data. As for the discrimination power of the items it is enough to know that every item is discriminating satisfactorily, you do not need to predict the exact level of discrimination.

In this study where the pretesting had been performed on large and representative samples it does not seem to be of any importance for the final test whether the item analysis has been performed within the IRT framework or within the CTT framework.

What is usually mentioned as the main shortcoming of CTT is that item statistics such as item difficulty and item discrimination depend on the particular examinee sample in which they are obtained, while this is not the case for IRT. The invariance of item parameters across groups is also claimed to be one of the most important characteristics of item response theory. For the authentic examinee groups used in this study it is difficult to find any obvious advantage or greater invariance in the IRT based item statistics.

International news

Professor Michal Beller reported that the Israeli lottery system for deciding topics for matriculation exams has now been abolished. A new reform is now on its way.

There has been strong and unscholarly propaganda lately against PET, carried out loudly by two faculty members from the School of Education in Tel-Aviv University. The test has been accused of being biased (mainly against poor people who cannot afford the coaching schools), and of being unvalid (even though extensive meta-analytic studies have shown the predictive validity to be between .40 - .50), etc.

NITE has also constituted an International Advisory Board, in which Michal Beller and Ronald Hambleton from this Board are members. Other members are Robert Brennan and Henry Braun and three Israeli members. Some of the main issues which have been discussed so far are: The need for selection. Advantages and disadvantages of tests. The net contribution of adding PET over using the matriculation certificate as a single predictor.

In Israel the Universities also have an informal, joint committee of representatives from all universities, that meets often with NITE to discuss policy issues for the PET.

Dr Linda Cook mentioned that a major effort for ETS is the computerization of the GRE, GMAT, and TOEFL. Although the computerization effort has overall been a successful one, some of the issues facing particularly the GRE, have been associated with the high cost of continual testing and the accessibility of testing sites. ETS has dealt with the issues of accessible testing sites for the GRE by establishing ”mobile” test sites in hotels, schools etc. TOEFL computerization will occur in July. Linda will update the Board about this effort when we meet next June.

Linda discussed issues related to the fairness of the SAT for subgroups, particularly women. She mentioned that some states in the US, Texas and California, are considering or had passed legislation abolishing affirmative action in admissions decision.

Professor Ronald Hambleton gave a short presentation of some research activities in the United States. He mentioned that most if not all credentialing agencies in the US were eager to have their exams administered via computer. Besides the obvious delivery problems of these exams on computers, measurement specialists were determining the size of item banks that might be needed to support computer based testing. Clearly, larger item banks will be needed and research is being carried out to determine strategies for expanding item banks. The use of more item writers, the use of item shells, the use of item alghoritms,

the use of superficial changes in items, etc. are being investigated. Computer based test designs are also being studied – everything from a single form of an exam being delivered to all candidates via computer to a fully adaptive computer based exam. Professor Hambleton also highlighted some of the research to develop new item formats to capitalize on the capability of the computer.

Professor Hambleton went on to say a few words about three additional topics test adaptions (translation) for use in foreign languages, new methods for standard setting (not yet validated), and estimation of item difficulties.

Large samples of examinees for field-testing new items are desirable, but they are also associated with high costs and a potential loss of item security. A potential solution to this dilemma might be to use test specialists to estimate item statistics. A study had been performed with the purpose to develop and field-test two methods for panelists to use in estimating the difficulties of LSAT items. One method was based on anchor descriptions and the other was based on item mapping to define the p-value scale. Three field-tests were carried out which revealed that there was still much to learn about the process of training panelists but at the same time some of the results were encouraging. Panelists indicated that they thought they could be trained to complete the estimation process with accuracy; they also demonstrated that they would benefit from discussion.