Constructing a Computerized Psychometric Adaptive Test for University Applicants with Disabilities

AVITAL MOSHINSKY and CATHRAEL KAZIN

NATIONAL INSTITUTE FOR TESTING & EVALUATION

JERUSALEM, ISRAEL

A Paper Presented at the Annual Meeting of the American Educational Research Association, New Orleans 2002

Abstract

In recent years, there has been a large increase in the number of university applicants requesting special accommodations for university entrance exams.

The Israeli National Institute for Testing and Evaluation (NITE) administers a Psychometric Entrance Test to assist universities in Israel in selecting undergraduates (comparable to the SAT in the USA). Since universities in Israel do not permit "flagging" of candidates receiving special testing accommodations, such scores are treated as identical to scores attained under regular testing conditions.

The increase in the number of students receiving testing accommodations and the prohibition on “flagging” have brought to the focus certain psychometric issues pertaining to the fairness of testing students with disabilities and the comparability of special and standard testing conditions.

To address these issues, NITE has developed a computerized adaptive psychometric test for administration to examinees with disabilities.

The paper discusses the process of developing the computerized test and ensuring its comparability to the P&P test. The paper also presents data on the operational computerized test.

Constructing a Computerized Psychometric Adaptive Test for University Applicants with Disabilities

Over the last few years, there has been a large increase in the number of university applicants requesting special testing accommodations for university entrance exams (Dana & Ziomek, 1996; Camara, Copeland & Rothschild, 1998). This increase has brought to the focus certain psychometric issues pertaining to the fairness of testing students with disabilities and the comparability of special and standard testing conditions (Geisinger, 1994).

One issue is how to compare test results obtained under standard and special conditions. In order to enable examinees with disabilities to demonstrate their full potential, changes in testing conditions are often provided. For example, examinees with disabilities taking a paper-and-pencil (P&P) test may receive such accommodations as additional time for the entire test or for specific sections, rest breaks between sections, or large print test booklets and answer sheets. However, providing such accommodations raises questions regarding the validity and fairness of the test. Once changes are made, the scores of two groups that were tested under different conditions might be no longer comparable. As a result, scores may either under-predict or over-predict academic performance and may advantage one group (whether examinees with disabilities or without disabilities) at the expense of the other (Sherman & Robinson, 1982, Braun, Ragosta & Kaplan, 1988; Wightman, 1993; Willingham, Ragosta, Bennett, Brown, Rock & Powers 1988).

The Solution – A Computerized Adaptive Test

The National Institute for Testing and Evaluation (NITE) in Israel administers a Psychometric Entrance Test to assist Israeli universities in selecting undergraduates (comparable to the SAT in the USA). To help address the problems of fairness and standardization, as well as accommodate the significant increase in the number of university

applicants requesting accommodated testing (whether for learning or physical disabilities), NITE has developed a computerized adaptive psychometric test.

Two fundamental considerations were involved in the development of this test. First, Israeli universities do not permit "flagging" of university candidates who take the admissions exam under special conditions; consequently, universities treat the score attained in these exams as a score attained in the regular exam (for further discussion of the problems related to the issue of "flagging," see Mandinach, Cahalan & Camara, 2002). Second, the computerized test was not designed to supplant the P&P test, but rather to be administered as an alternative for special examinees only. The computerized test was designated for individuals with disabilities; examinees without disabilities would continue to be tested using the P&P test. Therefore, it was necessary to validate, standardize and scale the computerized test so that its scores would be analogous and comparable to those of the standard test (Lord, 1980; Fan, Thompson & Davey, 1999; Thomasson, 1997).

The computerized test that was developed is based on the three-parameter logistic IRT model. (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan & Rogers, 1991)[1]. The computational capability of the computer makes it possible to administer an adaptive test in which questions are adapted to examinees' ability. Because the test if of an adaptive nature, examinees receive significantly fewer questions than in the P&P test (approximately half), making the test user-friendly and especially suitable for examinees with learning disabilities. This compensates for the time added to the test duration by the generous allotment of time per-item. The computerized test differs in some other significant respects from the P&P test. The characteristics of the computerized test -- such as enlarged fonts, separate presentation of

[1] This provides an excellent solution for the problems of standardization since the IRT model is characterized by: "item characteristics that are not group-dependent; scores describing examinees proficiency that are not test dependent; a model that does not require strictly parallel tests for assessing reliability." (Hambelton, Swaminathan & Rogers, 1991, page 5).

each item and time breaks between item types -- make it especially suitable for examinees with special needs, particularly those with learning disabilities.

Special Characteristics of the Computerized Test Developed at NITE

The NITECATSYS is a software package which implements computer adaptive testing based on the IRT model. The package consists of a number of programs that allow test generating, quality checking and the administration of a computerized test (Blum & Ronen, 2001). There are two additional modules that run with the NITECATSYS in administrations of the test. These modules are the Man-Machine Interface module and the administration module that handles the recording of examinees' personal details and scores[2] (Zach & Rahamim, 1999).

The most significant respect in which the computerized test differs from the P&P test as well as from other computerized tests is that time is allotted per-item and not per-section or per sub-test. This feature permits standardization at the level of a single item. In contrast, when time is allotted per section, examinees decide for themselves how much time to invest in each item (for example, spending more time on specific questions they find difficult, or dividing time among the various items because of test wiseness considerations). Consequently, large variability in examinees' performance time results. Another problem that occurs when time is allotted for a set of items, rather than per-item, is that not all examinees manage to finish the whole set of items, and they guess the answers for the last questions. As a result, their score is biased and inaccurate. The generous amount of time allotted per-item was determined empirically in several experiments according to time consumption at the level of the item-type (Rapp, Ronen & Cohen, 1996). The time allotted per-item and the total

[2] The MMI program is developed on a continuous basis. The module enables partial or complete audio reading of the test (an option that helps examinees with severe dyslexia) and, in the near future, will enable the enlargement of the text display (an option that helps examinees with severe vision impairment).

amount of time allotted per-section are detailed in Table 1. Also detailed in the table is the amount of time allotted per-section in the P&P test. In the computerized test examinees get about half the number of items and extended time allotment of about 100% - 400% compared to the to the P&P test[3]. These characteristics make the test non-speeded and especially suitable for examinees with disabilities (Donlon, 1973).

Table 1 about here

In sum, using CAT provides a number of meaningful advantages in a standardized "package deal": additional time allotment per item, fewer items, and a user-friendly Man-Machine Interface which includes large font, separate presentation of each item, and rest breaks. Though some of these accommodations are also available in the P&P modality, the computerized adaptive administration makes all these accommodations available simultaneously in a convenient and standardized way both for examinees and for test administrators. The reason for that is that in a computerized test administration, each examinee works according to it's own pace, unlike in a group administration of a P&P test, and therefore breaks and test beginning and ending are flexible. Also, a computerized test in which each item is presented separately in a large font size is less intimidating than a thick booklet which contains many pages of items.

The computerized adaptive test not only provides a standardized solution for providing accommodations, it also solves the even more problematic issue of ensuring fairness and comparability between tests administered with different time allotments .

[3] The mean time allotment per-item for the Verbal Reasoning section is ~2 minutes in the computerized test and 0.8 minutes in the P&P test; for the Quantitative Reasoning section, ~4 minutes in the computerized test and 1 minute in the P&P test; and for the English section ~3.5 minutes in the computerized test and 0.9 minutes in the P&P test).

The issue of extended time is the most significant source of noncomparability (Bridgeman & Cline, 2000; Willingham, 1997). It has been found that providing additional time affects the score (Camara, Copeland & Rothschild, 1998), though perhaps not as much for students without disabilities (Bridgeman, Trapani &, Curely 2002). Unidimensional models used with CATs implicitly assume that only knowledge of the correct answer, and not speed, is needed for estimating examinees' abilities (Hambleton & Swaminathan, 1985). Thus, CAT supplies an optimal solution for the extended time issue by providing a theoretical model in which there are no time constraints.

The advantages of a computerized "Universal Test Design" which suits the needs of all examinees both with and without disabilities has been emphasized previously (Bennett, 1995; Dolan & Hall, 2001). The goal of Universal Test Design is to anticipate from the outset what will work for people with different needs and build that into the design (rather than to add on "accommodations" for people with disabilities after the fact). This paper will describe the development of a CAT for examinees with disabilities which was constructed according to this "Universal Test Design" principle[4].

This paper describes the process of test development in three stages:

1. Developing a valid computerized IRT based test;
2. Validating the comparability of the computerized test to the P&P test;
3. Ensuring the suitability of the computerized test for examinees with disabilities.

[4] Though the CAT is suitable for examinees with or without disabilities, it was designed to parallel the regular P&P test and not to replace it.

Test Structure

The Psychometric Entrance Test tests three areas: Verbal Reasoning, Quantitative Reasoning and English as a Foreign Language. The computerized exam tests the same areas, and its items are selected from the standard P&P test. The three sections appear in a fixed order, as do the item-types within each section. The test contains the same types of items as the P&P test. The first two items in each section have an average level of difficulty and low discrimination. These items are sampled randomly from a unit of 6 items with similar characteristics. The rationale is that the first two questions should be neither too difficult nor too easy, and that the posterior variance (the error estimation) should not decrease too much after the first two items. There is a minimum number of questions and a maximum number of questions that each examinee gets. The test reaches its end once an examinee satisfies the criteria for finishing the test: 1. the posterior variance is smaller than the defined value, and 2. the examinee has finished the minimum number of questions as defined, or reached the maximum number of items. Test structure and the time allotted to each item-type are detailed in Table 1.

The test software enables various item-sampling rules, such as maximum information, difficulty of the item, a combination of maximum information and the difficulty of the item, random sampling and serial sampling. The usage of different sampling rules for different units controls the exposure of items and thus provides good psychometric indexes of the test.

Content equivalency

Before the computerized test was constructed, the use of a new, flexible test structure was considered. Such a structure would have a maximal bank and minimal limitations. However, it soon appeared that the content validity of such a test would be problematic, and

might cause examinees to perceive the test as invalid. Thus, the decision was made to build the computerized test with the same content limitations as the P&P test[5].

Item exposure

Item exposure is one of the major problems of computerized adaptive tests. Two solutions were designed to address the problem. First, the computerized test was designated for special examinees only, with examinees without disabilities continuing to be tested using the P&P test. Second, parallel versions of the test were constructed with relatively small pools of item-banks (about 150 items per section). A different version is used each time the test is administered. The various sampling rules as well as test simulations assisted in controlling item exposure. The mean of item exposure was 0.16 (SD=0.13).

Statistical Checks During the Development Process

Unidimensionality

The first step in the process was to develop a computerized adaptive test based on the three-logistic IRT model that fulfills the model's assumption. Unidimensionality was investigated for each of the three areas: Verbal Reasoning (Kaplan-Sheffer, Ben-Simon & Cohen, 1992), Quantitative Reasoning (Tractinsky, Ben-Simon & Cohen, 1989) and English (Ben-Simon, Tractinsky & Cohen, 1989). The procedure was based on the method developed by Rosenbaum (1984). It was found that the three sections fulfill the requirements of the model. Parameters were estimated by means of NITEST (Cohen and Budner, 1989) – a computer program for estimating IRT parameters under the three-parameter logistic model, similar to the ASCAL (Assessment Systems Corporation, 1987). Parameter estimation was

[5] For example, test simulations showed that examinees with low quantitative reasoning ability received four or five geometry questions about angles and triangles, and examinees with high quantitative reasoning ability got four or five questions of algebraic equations in which digits were switched with letters. Such questions appear often as tough "brainteasers" in newspapers. This flexible sampling rule made the test unbalanced, and might causing examinees to perceive it as invalid. After the structure of the CAT version was changed to incorporate similar content limitations as the P&P test, the exam looked quite similar to the P&P test

based on operational data of the P&P test. Calibration between P&P scores and computerized scores is based on previous experiments (Ben-Simon, Sheffer, Ronen & Cohen, 1993).

Test Simulations

Test simulations were conducted on 1500 examinees with a normal distribution of ability[6], in order to check the quality of the test during its development and to correct test structure according to the results of the simulations. Test simulations make it possible to satisfy psychometric requirements while minimizing item exposure. The more the test structure is controlled (e.g., in terms of item content, item-types, and sampling rules), the larger the intervention in the pure Maximum Information sampling rule, thus affecting the psychometric measurements. The less the test structure is controlled, the greater the potential to affect face validity and the greater the danger of item-exposure.

Results of the simulations (Table 2) showed that the mean posterior variance and validity (correlation between true and estimated theta) of the computerized test is similar to that of the P&P test. The error of the computerized test is smaller than the error of the P&P test on the edges of ability – that is, for examinees with low ability and for examinees with high ability. The error of the P&P test is lower for examinees with average ability (see figure 1 and figure 2). This finding is not surprising, since the P&P test contains many questions of about the same difficulty level, whereas the computerized test questions are adapted to examinees' ability.

Table 2 about here

[6] The simulations program gets the true theta (a theoretical variable) of the simulees (synthetic examinees) as an input and calculates the estimated theta of each simulee according to the defined structure of the test and according to item parameters. The simulated P&P calculates the estimated theta of simulees based on a complete version of a previously administered operational P&P test and its item parameters.

Table 2 also shows the correlations between the true ability (a theoretical variable) of simulees and their estimated ability, which is calculated according to the adaptive IRT program and according to the defined test structure. The correlations were found to be satisfactory and very similar in both the computerized test and in the P&P test (Moshinsky & Ronen, 1998). The mean number of questions in the P&P test (Table 2) is slightly higher than half the mean number of questions in the computerized test. This is a result of the adaptive nature of the test: examinees receive questions adapted to their ability, and therefore fewer questions are needed in order to estimate examinees' ability.

Assessing the comparability of the computerized test: Empirical evidence

Test simulations showed that the computerized test had satisfactory psychometric measurements compared to the P&P test by means of posterior variance and the correlation between the true and estimated theta.

One concern was that the modifications that were made to the test in the process of computerizing it (the different modality, the different measurement methods, different test structure, time allotment etc.) might affect its comparability to the P&P test. Those modifications might cause negative bias towards people who do not use computers[7], or towards other sub-groups (men vs. women). Although the simulations provided important pieces of information, they could not address such considerations as the user-friendliness of the program, the ability of examinees to concentrate for hours in front of a computer screen, and performance time. Some of these factors as well as others were also investigated by Cohen, Ben-Simon, Moshinsky & Eitan (2002). All differences between the computerized

[7] The computerized test was constructed in a way that required no previous experience in computer usage (for example, examinees do not use the mouse). This decision was made on the basis of surveys showing that one-third of all examinees do not use computers.

and P&P tests must be explored in order to ensure the validity of the computerized test and to ensure its comparability to the P&P test.

Two experiments were conducted in order to investigate the comparability of the computerized test and the P&P test (Heller & Moshinsky, 1999; Moshinsky, 2000). The results of the two experiments were very similar, and therefore will be reported together.

Method

One month before the administration of the operational Psychometric Entrance Test, a random sampling of applicants who had registered for the test were offered the opportunity to participate in an experiment in return for an early estimation of their score on the operational test. About 33% (667 applicants) of those who received the invitation agreed to participate in the experiment. Three hundred and thirty eight (338) of the participants were assigned randomly to the computerized test, and 329 to the P&P test (subjects were not informed in advance of this division).

Results

The mean scores of the two groups were compared to one another and were also compared to the scores achieved a month later in the operational test. The results of the experiment showed that scores on the computerized test were similar to scores in both the experimental and operational P&P tests (Table 3).

Table 3 about here

No statistically significant difference was found between the computerized group and the P&P group both in the experiment (in the Verbal Reasoning t=1.43, df=665, Pr(t)=0.15; in

the Quantitative Reasoning t=1.26, df=665, Pr(t)=0.21; in the English section reasoning t=0.59, df=665, Pr(t)=0.56; in all three sections size effects of these comparisons according to Cohen (1988) are smaller than 0.2[8]) and in the operational test (all t values=0.00 whereas the mean scores of the two experimental groups are almost similar). Hence the computerized test is comparable to the P&P test and predicts scores as well as the P&P test.

The difference of the mean scores between the first test (the experimental) and the second test (the operational) was about 30 points. This difference was slightly less than the regular mean difference between test and re-test, which is about 49 points. The short period (one month) between the test and re-test probably accounts for this finding; the minimal period for taking another test is usually six months.

The correlations between the scores in the experiment and the operational test appear in Table 4. These correlations are similar to the correlations usually found between test and re-test in the general population (Stoller & Allalouf, 2002), and they are even slightly higher, perhaps because of the short time gap between test and re-test in the experiment.

Table 4 about here

Two additional variables that might affect scores were examined: gender and previous experience with computers (Russell, 1999). The results show that these factors did not affect examinees' scores. In an analysis of variance in which the variables of modality (computer, P&P), gender (male, female) and computer usage (uses computer, does not use computer),

[8] Cohen's index (1988) of the size of the effect as a measurement for the size of the differences between the averages

the interaction between modality and gender was found to be insignificant (F=0.86, Pr(F)=0.35, DF=7,325); and the interaction between modality and computer use was also found to be insignificant (F=0.51, Pr(F)=0.48, DF=7,325).

In feedback questionnaires, subjects reported satisfaction with the computerized test and rated it as clear and user-friendly (Mean=3.8 on a scale of 1 to 4, SD=0.6). They also judged the test to be as fair and not more difficult than the P&P test (Mean=3.1 on a scale of 1 to 4, SD=0.8).

The comparative analyses that were performed on the computerized test and the P&P test, both in simulation and under experimental conditions, provided satisfactory data in regard to examinees' scores and other psychometric variables (such as test reliability and the posterior variance of the test).

Computers and Students with Disabilities: Past Research

Since 1985, when the revolution in personal computers began, many types of educational software have been developed. Research into the differences between P&P and computerized versions of tests and inventories has usually found no significant differences between them (for example, Kerchner & Kistinger, 1984; McDermott & Watkins, 1984; Horton & Lovitt, 1994). Researchers have also explored whether educational software can create differences in the scores of examinees without disabilities and examinees with disabilities due to different modalities of administration. Most researchers found that the different modality does not create differences between the two groups, and that in most cases the software improves the achievements of both examinees with and without disabilities (for example, Trifiletti, Frith, & Armstrong, 1984; Thorkildsen & Hansen, 1987; Burke, 1988;

Higgins & Boone, 1990; Engdahl, 1991; Tenny, 1992). However, some research has found that computer-administered dictation spelling tests may interfere with the cognitive processes required in spelling for students with disabilities or add to the problem-solving burden confronted by these students in conventional assessment with P&P tests (Varnhagen & Gerber, 1984; English, Gerber & Semmel, 1985).

Prior to constructing the computerized test, NITE surveyed students and found no significant differences in attitudes toward computers between students with and without learning disabilities (Moshinsky, Tenenbaum, Rapp & Ronen, 1997). The same finding was found also by Brown, Boscardin & Sireci (1999).

The Psychometric Entrance Test is a multiple-choice test which requires no spelling or writing. Thus, in light of the literature findings and the experiments that were conducted, demonstrating the comparability of the P&P test and the computerized test, the computerized test was found to be suitable for administering operationally to examinees with disabilities.

The Computerized Operational Test

The computerized adaptive test was first administered operationally in July 2000[9], 945 examinees who were entitled to testing with accommodations took the CAT version of the PET Of these examinees, 662 (70%) had been diagnosed as learning disabled (including ADD/ADHD)[10] and 283 (30%) had been diagnosed as having a physical disability. All these examinees need extra time in order to demonstrate their full potential. Examinees with severe learning disabilities (who must have the test read to them), examinees who are blind or have severely impaired vision continue to receive the P&P test with the appropriate

[9] Testing with accommodations is provided twice a year (in April and July). Each year, a different test version is administered.

[10] A special unit at NITE is responsible for evaluating examinees' anamnesis and learning disabilities diagnostic reports and determining which accommodation(s) are appropriate for examinees who request special testing accommodations.

accommodations. Examinees who need minor accommodations in the test format (such as large-print test booklet or answer sheet) also receive the P&P test.

In feedback questionnaires, examinees with disabilities reported satisfaction with the computerized test and rated it clear [?not sure what this means] and user-friendly (Mean=3.7 on a scale of 1 to 4, SD=0.6). They also judged the test to be generous in time allotment, compensating for slow reading and difficulties in concentration (Mean=3.5 on a scale of 1 to 4, SD=0.7) and they perceived the test to be fair (Mean=3.4 on a scale of 1 to 4, SD=0.9).

The computerized test was standardized and validated on examinees without disabilities, enabling a comparison of the various measurements of the test between examinees with and without with disabilities. Such a comparison had not been possible when examinees with disabilities got the same P&P test with extended time, since the tests (with and without accommodations) could not be scaled or standardized. However, scaling the computerized test (which featured such characteristics as generous time allotment per item) to the P&P test made it possible to compare the performance of examinees without disabilities and examinees with disabilities. Once the computerized test became operational, the performance time, scores and quality of responses to different types of items of examinees without disabilities (who participated in the above experiment) and examinees with disabilities could be compared. These data can be instructive regarding the relative performance of examinees with disabilities and examinees without disabilities with respect to reading rate and specific mental tasks in which difficulties are expressed, as well as cognitive tasks in which examinees with disabilities who have received extended time, perform as well as examinees without disabilities. The following analyses were made:

1. A comparison between scores of examinees without disabilities who were tested in the experimental computerized test and examinees with physical disabilities or learning disabilities who took the same test in an operational administration.It is

important to remember that these groups differ not only with respect to their learning abilities or disabilities, but also in terms of their motivation in performing the test, since one group took the test as an operational test and the other in the context of an experiment.

2. A hypothetical comparison, based on the past scores of examinees with disabilities who would have been eligible to be tested with the computerized test had it existed at the time, and the current scores of examinees with disabilities who were tested in the computerized test. This can give important (if necessarily somewhat problematic) information that is not anchored to examinees without disabilities.

3. A comparison of the actual scores of examinees with disabilities on the computerized test and on previous P&P tests (with or without special accommodations). This group is not fully representative of the total group of examinees with disabilities, but nevertheless the comparison provides significant additional information.

Even though these comparisons are somewhat problematic because they compare different populations that were tested in different years in different modalities, they supply important information about the computerized test.

Table 5 about here

The results of the first comparison appear in Table 5. According to the Table, there is a difference in the mean scores of examinees with learning disabilities and examinees without learning disabilities (either without disabilities or with physical disabilities). However, this difference is not due to different modality, since there is no difference in the mean scores of examinees who were tested in the computerized test compared to examinees who were tested

with accommodations in the P&P test prior to the development of the computerized test (F=0.10, Pr(F)=0.75, DF=3,2785) ; the size of the effect according to Cohen's index (1988) is negligible (0.02 std units).

Results of the comparison between examinees without disabilities and examinees with physical disabilities show that the mean score in Verbal Reasoning and Quantitative Reasoning of examinees with physical disabilities is equal to the mean score of examinees without disabilities, and the difference in the mean scores of the English section is slightly lower for examinees with physical disabilities t=2.63, df=1135, Pr(t)=0.004; the size of the effect according to Cohen's index (1988) is small (0.17 std units).

It is also apparent that the mean scores in all three sections of examinees with learning disabilities are lower than the mean scores of examinees without learning disabilities (either without disabilities or with physical disabilities). This difference is significant (F=30.80, Pr(F)<.0001, DF=3,2785) ; the size of the effect according to Cohen's index (1988) is small (0.26 std units). These differences are either a result of ability differences, or disabilities that cannot be compensated by the given accommodationa (for example, acquiring a foreign language is more difficult for examinees with disabilities (Sparks, Granschow & Javorsky, 1992)). Results of the comparison between scores differences of examinees with disabilities who were tested in the past with or without special accommodations (between test and re-test) are presented in Table 6. No significant difference was found between the mean difference of test and re-test of examinees with learning disabilities and examinees without learning disabilities and therefore the results in Table 6 are reported for these groups together.

Table 6 about here

Results show that the mean improvement in the scores of examinees who were tested in the first test with the regular P&P test and in the second test with the computerized test is 88 points, which is 39 points higher than the average 49 point improvement between test and re-test among the general population (Stoller & Allalouf, 2002)[11]. The average improvement among examinees who were tested in the past with the P&P test with accommodations and in the computerized test was 58 points. An above-average improvement was also seen between the test and re-test of examinees who were first tested with the regular P&P test (without accommodations) and later with the P&P test with accommodations: the mean difference was 63 points, 14 points above the average improvement. An approximate average improvement of 40 points is seen among examinees which were tested both in test and re-test in the P&P test with accommodations. These findings are understandable, because the accommodations are intended to help examinees to demonstrate their true potential.

An analysis of variance was conducted between the modality in the first test (regular P&P test, P&P test with accommodations) and the modality in the second test (computerized test, P&P with accommodations); the dependent variable was the score difference between test and re-test. The first modality was found to be significant, i.e., the improvement was higher for those who were tested in the first test without accommodations than for those who were tested with accommodations ($F=21.13$, $Pr(F)<.0001$, $DF=3,474$); the size of the effect according to Cohen's index (1988) is medium (0.56 std units). The second modality was also found to be significant, i.e., the improvement was higher for those who were tested the second time with the computerized test ($F=13.37$, $Pr(F)=0.0003$, $DF=3,474$); the size of the effect according to Cohen's index (1988) is medium (0.42 std units).

[11] The mean improvement between test and re-test in the years 1995-2000 among 68,774 examinees was 9 points (Std = 11) in the Verbal Reasoning section, 8 point (Std = 10) in the quantitative reasoning section and 9 points (Std = 11) in the English section (Stoller & Allalouf, 2002).

The interaction between the first modality and the second modality was not significant (F=0.30, Pr(F)=0.59, DF=3,474). Similar results were accepted for the mean differences in each of the three sections.

The finding may have limited significance, however, because this group is non-random. The examines who were tested with the P&P with accommodations have different disabilities than those who were tested with the computerized test (e.g., they have severe dyslexia and received an audio version of the test or have severe vision limitations).

Thus, Table 6 suggests that the computerized test enables examinees with disabilities to demonstrate their potential as much as, and even more than, the P&P test with accommodations.

Analysis of the computerized test's data also shows that performance time of examinees with disabilities is significantly longer than performance time of examinees without disabilities for all three areas of the test and for all item-types (Table 7).

Table 7 about here

Though the mean time duration of the test is significantly different for the two groups of examinees, the proportions of time allotted for each section are quite similar. Examinees with disabilities seem to need more time breaks, probably due to the extended time (an additional hour) it takes them to complete the exam.

Further research regarding these data and regarding performance time of each item-type is being conducted.

Implications

The computerized test provides a solution for a number of psychometric problems, offering:

1. A standardized administration of the test for examinees with disabilities.
2. The ability to compare scores of examinees without disabilities and examinees with disabilities.
3. Facilitation of the process of decision-making, since it is no longer necessary to decide which examinee will be granted which of the many conditions and time additions that are given in the P&P test.
4. A solution for examinees with disabilities by providing them with several accommodations in a standardized "package" that can be offered both to examinees with and without disabilities in accordance with the concept of "universal test design".

In summary, the computerized test provides a satisfactory solution both for the psychometric problems presented by the increasing number of university applicants with disabilities and for the special needs of examinees with disabilities.

Further research regarding the predictive validity of the computerized test (in particular, the correlation between test scores and university academic record) is still needed.

Acknowledgements

We wish to thank **Naomi Gafni** for her advice, support and helpful comments throughout the entire process of conducting the project and compiling this paper.

We wish to thank the members of the CAT unit –
Nadav Blum, Ilana Rahamim and Boaz Maori for their dedicated efforts during the implementation of this project.

We would also like to thank **Yigal Attali** for his helpful comments.

References

Assessment Systems Corporation, (1987). *User's Manual for MicroCAT Testing System*. 2nd Ed. St. Paul, Minnesota.

Ben-Simon, A., Tractinsky, N. & Cohen, Y. (1989). Item-Banking of EFL Items using the 3-p logistic model. *NITE, Report #103*, Jerusalem, Israel.

Ben-Simon, A., Sheffer, L., Ronen, T. & Cohen, Y. (1993). A computerized adaptive version of the PET for self evaluation of ability level. *NITE, Report #175*, Jerusalem, Israel.

Bennett, R.E. (1995). Computer-based testing for examinees with disabilities: On the road to generalized accommodation. In S. Messick (Ed.), *Assessment in higher education: Issues of access, student development, and public policy*. Hillsdale, NJ: Erlbaum.

Blum, N. & Ronen, T. (2001). A technical review of the operation module in the NITECATSYS computerized test package. *NITE, Report*, Jerusalem, Israel.

Braun, H., Ragosta, M. & Kaplan, B. (1988). *Predictive validity.* In Willingham, W., Ragosta, M., Bennett, R. E., Braun, H., Rock, D. A. & Powers, D. E., *Testing handicapped people.* Boston: Allyn and Bacon.

Bridgeman, B, & Cline, F. (2000). *Variations in mean response times for questions on the computer-adaptive GRE General Test: Implications for fair assess*ment (GRE Board Professional Report No. 96-20P; ETS RR 00-7).

Bridgeman, B., Trapani, C. & Curley, E. (2002). Effect of Fewer Questions per Section on SAT I Scores. *College Board Research Report No. 2003-2 ETS RR-03-08*

Brown, C. R. Boscardin, M. L. & Sireci, S. G. (2001). *Computer Attitudes and Opinions of Students with and without Learning Disabilities. Journal of Educational Computing Research*, 24, 183-204.

Burke, J. P. (1988). *Improving the Perceptual Performance of Learning Disabled Second Graders through Computer Assisted Instruction.* Reports Descriptive (141), Florida, U.S.

Camara, W.J., Copeland, T. & Rothschild, B. (1998). Effects of extended time on the SAT I: Reasoning test score growth for students with learning disabilities. *College Board Report No. 98-7.*

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd edition). Hillsdale, NJ: Erlbaum.

Cohen, Y. & Budner, G. (1989). A manual for NITEST – a program for estimating IRT parameters. *NITE, Report #94*, Jerusalem, Israel.

Cohen, Y., Ben-Simon, A., Moshinsky, A. & Eitan, M. (2002). *Computer based testing (CBT) in the service of test accommodations.* Paper presented at the annual meeting of the International Association for Educational Assessment, Hong Kong.

Dana, P. & Ziomek, R. L. (1996). *Extended time testingon ACT or SAT: What difference does it make?* Paper presented at the National Association of College Admissions Counsellers, Minneapolis, MN.

Dolan, R. P. & Hall, T. E. (2001). Universal Design for Learning: Implications for Large-Scale Assessment, *IDA* Perspective,s 27(4): 22-25.

Donlon, T. F. (1973). *Establishing appropriate time limits for tests*. Paper presented at the fall meeting of the Northeast Educational Research Association, Ellenvile, NY.

Engdahl, B. (1991). *Computerized Adaptive Assessment of Cognitive Abilities among Disabled Adults*. Paper presented at the Annual Meeting of the American Psychological Association (San Francisco, CA, August 16-20, 1991).

English, J. P., Gerber, M. M. & Semmel, M. I. (1985). Microcomputer administered spelling tests: Effects on learning handicapped and normally achieving students. *Journal of Reading, Writing and Learning Disabilities International*. Vol 1(2): 165-176.

Fan, M., Thompson, T. & Davey, T. (1999). *Constructing adaptive tests to parallel conventional programs*. Paper presented at the National Council of Measurement in Education meeting, Montreal, Canada, April.

Geisinger, K. F. (1994). Psychometric issues in testing students with disabilities. *Applied Measurement in Education*, 7(2), 121-140.

Hambleton, R. K. & Swaminathan, H. (1985). *Item Response Theory, Principles and Applications*. Kluwer, Nijhoff Publishing: Boston.

Hambleton, R. K., Swaminathan, H. & Rogers, H. J. (1991*). Fundamentals of Item Response Theory*. Sage Publications: Newbury Park London New Delhi.

Heller, D. & Moshinsky, A. (1998). Computerized Adaptive Test for the examinees with disabilities – version 1: Results of Experiment 1. *NITE, Report #256*, Jerusalem, Israel.

Higgins, K. & Boone, R. (1990). Hypertext computer study guides and the social studies achievement of students with learning disabilities, remedial students, and regular education students. *Journal of Learning Disabilities*. Vol 23(9): 529-540.

Horton, S. V. & Lovitt, T. C. (1994). A comparison of two methods of administering group reading inventories to diverse learners: Computer versus pencil and paper. *Remedial and Special Education*. Vol 15(6): 378-390

Kaplan-Sheffer, L., Ben-Simon, A. & Cohen, Y. (1992). Item-Banking of Verbal Reasoning Items using the 3-p logistic model. *NITE, Report #165*, Jerusalem, Israel.

Lord, F. M. (1980). Applications of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Kerchner, L. B. & Kistinger, B. J. (1984). Language processing/word processing: Written expression, computers and learning disabled students. *Learning Disability Quarterly*. Vol 7(4): 329-335

Mandinach, E. B., Cahalan, C. & Camara, W.J. (2002). The Impact of Flagging on the Admission Process: Policies, Practices, and Implications. *College Board Research Report No. 2002-2*

McDermott, P. A. & Watkins, M. W. (1984). Computerized vs. conventional remedial instruction for learning-disabled pupils. *Journal of Special Education.* Vol 17(1): 81-88.

Moshinsky, A. (2000). Computerized Adaptive Test for examinees with disabilities – version 2: Results of Experiment 2. *NITE, Report #276,* Jerusalem, Israel.

Moshinsky, A. & Ronen, T. (1998). Computerized Adaptive Test for examinees with disabilities – Versions 1,2: Test's Structure and Simulations' Results. *NITE, Technical Report #85,* Jerusalem, Israel.

Moshinsky, A., Tenenbaum, M., Rapp, J. & Ronen, T. (1997). The habits of using computers among examinees with learning disabilities and physically handicapped. *NITE, Technical Report #72,* Jerusalem, Israel.

Rapp, Y., Ronen, T. & Cohen, Y. (1996). Analyzing Item Type Performance time in an Experimental Administration of the CPETSE. *NITE, Technical Report #52a,* Jerusalem, Israel.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of Item Response Theory. *Psychometrika*, 49, 425-435.

Russell, M. (1999). *Testing on computers: A follow-up study comparing performance on computer and paper*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada, April.

Sherman, S. & Robinson, N. (Eds.). (1982). *Ability testing of handicapped people: Dilemma for government, science and the public.* Washington, DC: National

Sparks, R., Granschow, L. & Javorsky, J. (1992). Diagnosing and accommodating for foreign language learning difficulties of college students with learning disabilities. *Learning Disabilities Research & Practice*, 7: 150-160.

Stoller, R. & Allalouf, A. (2002). *Improvement in re-tests.* NITE Report (in preparation).

Tenny, J. L. (1992). Computer-Supported Study Strategies for Purple People. *Reading and Writing Quarterly: Overcoming Learning-Difficulties*, 359-77 Oct-Dec 1992.

Thomasson, G. L. (1997). *The goal of equity within and between Computerized Adaptive Tests and Paper and Pencil forms*. Paper presented at the National Council of Measurement in Education meeting, Chicago, IL, March.

Thorkildsen, R. J. & Hansen, P. (1987). *Development and Field Testing of a Videodisc/Teacher Net System for Mildly Handicapped Students. Final Report.* Reports Research (143), Utah, U.S.

Tractinsky, N., Ben-Simon, A. & Cohen, Y. (1989). Item-Banking of Quantitative Reasoning Items using the 3-p logistic model. *NITE, Report #90*, Jerusalem, Israel.

Trifiletti, J. J., Frith, G. H. & Armstrong, S. W. (1984). Microcomputers versus resource rooms for LD students: A preliminary investigation of the effects on math skills. *Learning Disability Quarterly*. Vol 7(1): 69-76.

Varnhagen, S. & Gerber, M. M. (1984). Use of microcomputers for spelling assessment: Reasons to be cautious. *Learning Disability Quarterly.* Vol 7(3): 266-270.

Wightman, L. F. (1993). Test takers with disabilities: A summary of data from special administrations of the LSAT. *Research Report 93-03. Law School Admission Council.* Newtown, PA: LSAT and Law School Admissions Services.

Willingham, W., Ragosta, M., Bennett, R. E., Braun, H., Rock, D. A. & Powers, D. E. (1988). *Testing handicapped people.* Boston: Allyn and Bacon.

Zach, Y. & Rahamim, I. (1999). NITECATSYS: Man-Machine Interface (MMI). *NITE, Technical Report #94,* Jerusalem, Israel.

Table 1

Computerized adaptive PET structure:
Test domains and item-types, time allotment per item-type, number of items of each item-type

Domain	Item-type	Time allotment per-item (in minutes)	Number of items in single test
Verbal Reasoning	Verbal Analogies	1.5	8-13
	Sentence Completion	2	6
	Logic	3	4
	Letter- Switching	3	4
	Words and Phrases	1	3-5
	Reading Comprehension	7 (per-text) 4 (per-item)	5
Total - Computerized	-	66-76 minutes	30-37
Total - P&P	-	50 minutes	60
Quantitative Reasoning	Algebra	4	13-18
	Geometry	4	5
	Quantitative Comparison	4	6
	Graph or table	5 (per-graph) 4 (per-item)	4
Total - Computerized	-	117-122 min.	27-32
Total - P&P	-	50 minutes	50
English	Sentence Completion	2	8-17
	Restatements	4	8
	Reading Comprehension	7 (per-text) 4 (per-item)	5
Total - Computerized	-	75-89 minutes	21-28
Total - P&P	-	50 minutes	54

Table 2

Results of test simulations:

Comparison between computerized and P&P test structures

Modality	Test Area	Mean number of questions (Std)	Correlation between true and estimated theta	Posterior Variance (Std)		
				$\theta<-1.5$ n=97	$-1.5<\theta<1.5$ n=1304	$1.5<\theta$ n=97
P&P test	Verbal Reasoning	60 (0)	0.94	0.29 (0.12)	0.13 (0.04)	0.16 (0.04)
	Quant. Reasoning	50 (0)	0.95	0.23 (0.07)	0.12 (0.03)	0.14 (0.04)
	English	54 (0)	0.96	0.21 (0.10)	0.06 (0.03)	0.15 (0.07)
Comp. test	Verbal Reasoning	36 (2)	0.92	0.17 (0.03)	0.14 (0.01)	0.15 (0.02)
	Quant. Reasoning	31 (2)	0.95	0.20 (0.05)	0.11 (0.02)	0.12 (0.03)
	English	23 (3)	0.96	0.16 (0.08)	0.07 (0.02)	0.09 (0.04)

Table 3

Mean scores* of computer and P&P groups in experimental and subsequent operational tests

	Experiment Scores (Std)		Operational Scores (Std)	
	Computer Group (n=338)	P&P Group (n=329)	Computer Group (n=338)	P&P Group (n=329)
Verbal Reasoning	111 (18)	113 (20)	116 (19)	116 (19)
Quant. Reasoning	113 (19)	111 (17)	118 (18)	118 (17)
English	113 (22)	114 (22)	116 (21)	117 (21)
Total	569 (96)	573 (94)	598 (98)	598 (94)

* The range of scores is 50-150 for each of the three areas, and 200-800 for the total score.

Table 4

Correlations between experimental (P&P and Computer Groups) and subsequent operational test scores, and test and retest in general population

	P&P Group *N=329*	*Computer Group* *N=339*	*General Population* *N=21,792*
Verbal Reasoning	.89	.86	.79
Quantitative Reasoning	.85	.84	.81
English	.93	.92	.86
Total score	.94	.92	.88

Table 5

Mean scores* of examinees without disabilities (computerized test), examinees with disabilities (computerized test), and examinees with disabilities (P&P test, who would have been eligible for computerized test had it been available)

	Examinees tested in the experiment N=338	Examinees tested in the computerized test (2000-2003) N=945		Examinees from previous years (1997-1999) tested in P&P test N=1844	
	without disabilities	Leaning disability n=662	Physical disability n=283	Leaning disability n=1328	Physical disability n=516
Verbal Reasoning	111 (18)	108 (18)	111 (20)	106 (18)	111 (18)
Quantitative Reasoning	113 (19)	110 (21)	113 (23)	108 (19)	113 (19)
English	113 (22)	102 (24)	107 (27)	106 (22)	110 (22)
Total Score	569 (96)	543 (94)	562 (114)	539 (92)	566 (92)

* The range of scores is 50-150 for each of the three areas, and 200-800 for the total score.

Table 6

Mean scores* differences of examinees with disabilities who were tested in the past with or without special accommodations between test and re-test for the years 2000-2003

	Test I – regular P&P		Test I – P&P with accommodations	
	Test II – computerized n=185	Test II – P&P with accom. n=181	Test II – computerized n=54	Test II – P&P with accom. n=57
Mean difference of Verbal Reasoning	17 (14)	12 (13)	12 (13)	9 (11)
Mean difference of Quantitative Reasoning	14 (12)	10 (11)	10 (13)	5 (10)
Mean difference of English	15 (16)	12 (11)	7 (13)	9 (11)
Mean difference of Total Score	88 (60)	63 (52)	58 (52)	40 (39)

* The range of scores is 50-150 for each of the three areas, and 200-800 for the total score.

Table 7

Mean time duration of the test and percentage of time devoted to each of the three sections by examinees with and without disabilities

	Examinees without disabilities *N=338*	*Examinees with disabilities* *N=944*
Mean time duration (in minutes) of the test	141, SD=34 Min=76, Max=225	199, SD=47 Min=68, Max=339
% of time devoted to Verbal Reasoning	31%	33%
% of time devoted to Quantitative Reasoning	39%	35%
% of time devoted to English section	22%	22%
% of time devoted to instructions and breaks	8%	10%

Figure 1

Simulations results (Quantitative Reasoning) – a comparison of the posterior variance as a function of theta in the P&P test and in the computerized test

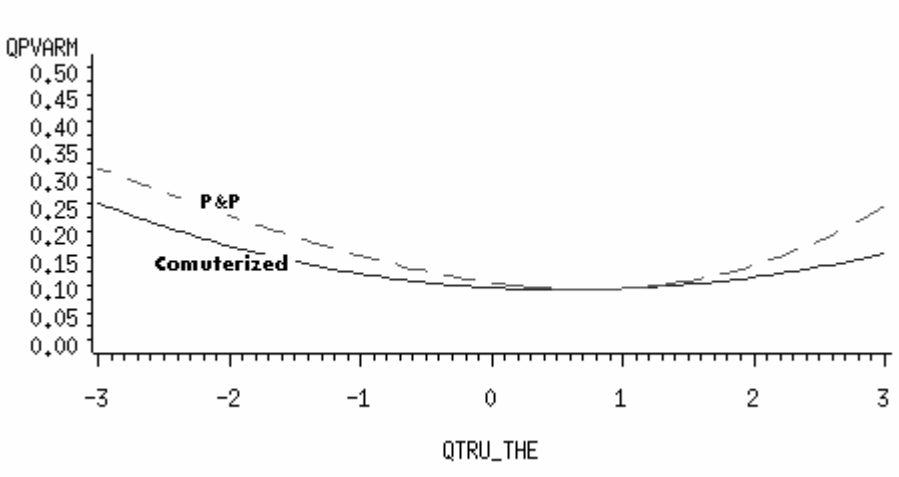

Figure 2

Distribution of the posterior variance at the various ability levels in the English section