דוח המרכז

348

Computer Based Testing (CBT) in the Service of Test Accommodations

Yoav Cohen
Anat Ben-Simon
Avital Moshinsky
Miriam Eitan

July 2008

מרכז ארצי לבחינות ולהערכה (ע"ר)
NATIONAL INSTITUTE FOR TESTING & EVALUATION
מיסודן של האוניברסיטאות בישראל
ת"ד 26015 ירושלים 91260, טל' 02-6759555

דוח מרכז 348

ISBN:965-502-141-6

Computer Based Testing (CBT) in the Service of Test Accommodations

Yoav Cohen, Anat Ben-Simon, Avital Moshinsky and Miriam Eitan

National Institute for Testing & Evaluation (NITE), Jerusalem, Israel

Abstract In the last two decades there has been an increase in the number of university applicants who are diagnosed as learning disabled (LD) and for whom test accommodations on university entrance exams are provided. The most frequent recommendation in the diagnostic reports of LD applicants is to extend the time limits of their tests. In the context of high-stakes testing, this kind of accommodation raises the question of equity: is it fair to extend the time limit of a speeded test to a particular group of examinees? Does it really give the LD a fair chance? And if so, by how much should the time limit be extended? Administering a computer-based version of the test to the LD can largely circumvent these issues.

University applicants in Israel were required, until recently, to submit scores on the Psychometric Entrance Test (PET) to universities. This paper discusses the issues associated with test accommodations in general and with PET accommodations in particular. It then describes the process of constructing a Computerized Adaptive Test (CAT) which is equivalent to the paper and pencil version of the PET, and presents data pertaining to the equivalence of the CAT and paper & pencil versions of the test.

In the last two decades there has been an increase in the number of university candidates who are diagnosed as learning disabled (LD) and for whom test accommodations on university entrance exams are provided. The most frequent recommendation in the diagnostic reports of LD candidates is to extend the time limits of their tests. In the context of high-stakes testing, this kind of accommodation raises the question of equity: is it fair to extend the time limit of a speeded test to a particular group of examinees? Does it really give the LD a fair chance? And if so, by how much should the time limit be extended? The answers to these questions are not clear-cut. Full answers require extensive research and even with good research results, their utility depends, to large extent, on the quality and fidelity of the LD diagnostic report.

In this paper, we would like to discuss the application of computer-based testing that, in our opinion, circumvents these problems and yields a satisfactory solution. More specifically, the kind of computer based test (CBT) that we have adopted for testing the learning disabled is a computerized adaptive test (CAT). We will first describe

CAT and discuss its main characteristics and advantages, then discuss the problems associated with test accommodations. We then describe the research projects that were undertaken in order to develop a computerized version of a high stakes test and lastly, report on the actual implementation of the CAT for the learning disabled.

Computerized adaptive tests

Adaptive tests are tests in which the items are selected such that they best suit the knowledge or ability level of the examinee. Thus, in an adaptive test, a high level examinee will not be asked to respond to extremely easy items, and a low level examinee will not be confronted with items whose difficulty is such that the probability of answering them correctly (other than on the basis of sheer guesswork) is nil. So, in order to adapt the test one has to know something about the level of the examinee beforehand. This might sound like a vicious circle, because in order to know the level of the examinee you must test her, but in practice, there are ways around the problem. In individually administered intelligence tests, for example, each child usually sees only a small part of the items in each scale. The items on each scale are ordered from the easy to the difficult and the first item that is presented to the child is determined by her age. If the child misses the first few items, then easier items are presented; testing continues until the child fails a pre-set number of consecutive items. In testing adults, or in the case of achievement tests, age cannot serve as a good index for the general level of the examinee. In these cases, some form of trial and error is performed at the beginning of the test. Generally, the first item is sampled from a pool of easy to moderate items. The difficulty level of the next item is determined by the answer to the first one. An easier item will be provided if the examinee missed the first item, and a more difficult item will be provided if the examinee answered the first item correctly. Thus, after each set of items, the ability level of the examinee is estimated and the difficulty level of the next item is determined by this estimate.

The nature of the adaptive process is such that any two examinees do not necessarily answer the same items. In facts, two examinees might even be tested on mutually exclusive sets of items. From this description, it becomes evident that in addition to a mechanism for adapting the test, a strong measurement model is needed in order to help in scoring tests that are based on different sets of items.

Computerized adaptive testing was born out of the marriage of measurement models of Item Response Theory (IRT) (Birnbaum 1968, Lord 1980, Hambleton 1985) and the technology of personal computing which evolved so rapidly in the last two decades.

There are numerous psychometric and logistic advantages to computerized adaptive tests. From the psychometric point of view, the most important advantage of CAT is the high accuracy of the measurement that is achieved with a relatively small number of items. In one example (Cohen, Ben-Simon and Tractinsky, 1989), a computerized adaptive test of language proficiency achieved a test-retest reliability comparable to that of a Paper and Pencil (P&P) non-adaptive test with the use of only half as many items. The second advantage of CAT is its increased bandwidth, that is, the spread of ability levels that it can measure without sacrificing the fidelity of measurement. A third advantage of computerized tests is the ability of the tester to allocate time for each and every item, instead of allocating total time for a section of a test, and, on the other hand, the ability to measure response latency, which provides an additional source of information regarding the examinee. A fourth advantage of computerized tests, and probably the most promising, is the option of using new modes of stimuli and response for the measurement of behavior. From the logistic or administrative perspective, a further advantage of computerized tests is the fact that a scoring report is generated immediately and can be provided to the examinee as soon as the test terminates. A second advantage is the reduced time of testing, which is a direct result of the reduced number of items in a CAT relative to a P&P test.

The gains of CAT do not come without a price. In the case of CAT the price is in terms of items, psychometric constraints and the technological infrastructure that is needed to support this mode of testing. CATs are based on large item pools from which the testing algorithm samples the items that are actually presented to the examinee. Thus, for each test-form of a CAT, the number of items needed is at least twice that of a P&P test form, even though the number of items actually presented to the examinee is half of that of the P&P test. In addition, P&P tests are usually administered to large group of examinees in one sitting. This is in contrast to computerized tests that are administered in several sittings, each time to a relatively small group of examinees. Therefore, after several administrations of the CAT, and

especially when the tests are of high stakes, the items have to be retired since they can no longer assumed to be secure. This forces the testing agency to constantly update the item pools. The second price that has to be paid for CATs is in terms of psychometric constraints. Since the adaptive tests are based on IRT, which in turn, requires unidimensionality (or essential unidimensionality) of item pools, one has to ascertain that the item pools are indeed unidimensional, and, at the same time, cover the range of abilities or content areas that are covered by the P&P test.

Potential applications of CAT

CAT can be applied wherever a P&P is used (provided, of course, that the suitable hardware is available), but there are some contexts in which it can be more useful than in others. Using tests for placement purposes is one such context. The immediate feedback that is provided by computerized testing, and the wide bandwidth of adaptive testing make CATs ideal in situations in which fast decisions regarding a heterogeneous population are called for. Indeed, one of the first applications of CAT was in the testing of English as a foreign language, in the context of in-house placement of students. This was done by the Educational Testing Service in the US and by the National Institute for Testing and Evaluation (NITE) in Israel (Cohen et al. 1989). Another context in which the application of CAT's seems beneficial is of testing people with disabilities who need special testing accommodations. Of these, the largest group of examinees is that of people who are diagnosed as learning disabled (LD).

Special populations of examinees

A significant percentage of examinees, whether students, university applicants, or job candidates, are defined as individuals with disabilities who need and are entitled to special accommodations in testing. There is growing awareness of their rights and needs, and this is also recognized by the legislature. For example, a law has been proposed by the Israeli parliament that would ensure the rights of LD students in primary and secondary educational institutions. The population of people with special needs can be roughly divided into three groups: the physically disabled, the learning disabled and the emotionally or psychologically disabled. In addition to the

differences in the nature of their disabilities, these groups also differ in the nature of testing accommodations that are provided. Furthermore, they differ with respect to the type of their diagnoses. Physical problems are, in most cases, diagnosed and reported by physicians who use established techniques and tools. In the case of the learning disabled, the diagnosis is sometimes made by physicians and at other times by educational psychologists, clinical psychologists or by educational counselors who have received special training. In the case of LD diagnosis there is less agreement among practitioners regarding the models and techniques that are best suited for diagnosis and the diagnostic tools are not always standardized and objective (Eitan et. al, 2002). The diagnostic process is even more subjective in the case of psychological and emotional problems such as test anxiety. As discussed below, the extent to which the diagnosis is standardized and objective has some bearing upon the decision how to accommodate the test.

Test accommodations

The ways in which tests can be, and are, in practice, accommodated are quite numerous. Thus, for example, tests can be given in different modalities: either visual (e.g., large print), or auditory (read aloud) or tactile (in Braille); special seating conditions are often provided for people with physical disabilities, or those who need to be tested alone because they are easily distracted, and so forth. The full gamut of accommodations that are provided by NITE to examinees is listed in table 1, but among those who apply for special accommodations, the largest group is that of the LD who, in most cases, ask for extra testing time.

The questions to whom to grant accommodation, what is the proper basis for this decision, and exactly what kind of accommodation to give to each applicant are questions that directly relate to the validity of the test. Thus, for example, if a test in mathematics is administered without access to a calculator but a calculator is provided to some examinees as a means to compensate for discalculia, then one might ask whether providing a calculator to all examinees will not help them all. If the answer is in the affirmative, then by giving a calculator to the few we create a bias that hampers the validity of the test. On the other hand, if simple calculations are needed, but they are used in order to gain information about the reading ability of the

examinee, and not about her mathematical abilities, then depriving her of the calculator will lead to underestimation of her true score.

NITE has recently adopted the following general principles regarding test accommodations (Ben-Simon, 2001; Eitan, 2002). First, it is recognized that the need for accommodations stems from two sources, those that involve societal values such as equal rights and social justice, and those that involve professional/psychometric considerations such as accurate measurement of abilities. Second, the goal of providing test accommodations is to make sure that the measurement will not be affected by irrelevant factors. Third, one has to consider both the examinee and the nature of the test in order to determine whether, how, and to what degree, to provide test accommodations. One has to consider the exact nature of the examinee's disabilities and at the same time to consider the nature of the abilities that are measured by the test.

Extending the time limit of a test

As mentioned above, a significant proportion of those candidates requesting test accommodations have been diagnosed as LD. Of those diagnosed as learning disabled, the majority are classified as Reading Disabled; the most common recommendation by the LD specialists is to extend their testing time, usually by allotting them 20% to 50% more time than the time allotted for regular examinees. This recommendation raises three questions. First, will extending the time help those who are diagnosed as reading disabled and will it indeed compensate for their disability? Second, what would be the effect of time extension on the performance of the normal population? Third, how accurate, professional and standard is the determination of their reading disability?

Extending the time limit for the reading disabled indeed helps them in testing situations. They, and their diagnosticians, report that it takes them much longer to read texts, hence, in every speeded test that involves a significant amount of reading they benefit from extra testing time. So apparently, and according to subjective reports, extension of time limits compensates for the difficulties encountered by the reading disabled. This answers the first question. As for the second question –

according to some researchers (e.g. Runyan and Smith, 1991), though LD students benefit from extra time, non-learning disabled students would not benefit from extra time because they are already working at their maximum potential under timed conditions. But as Zuriff (2001) has shown, this is in contrast to his findings that non-learning disabled students may also benefit from extra examination time. Thus, the answer to the second question is that extra time may in fact hamper the validity of timed tests. As for the third question, according to an extensive review of diagnostic reports that were collected from files of university applicants at NITE (Eitan, Moshinsky & Ben-Simon, In preparation), it seems that the techniques, tools and standards in the diagnosis of reading disability are far from optimal. For example, it was found that the same level of reading fluency, as indexed by the number of words per minute, was classified as 'slow' by some diagnosticians, 'medium' by others, and 'fast' by the rest. In any event, if every person who is diagnosed as having reading disability receives extended testing time, we are left with the nearly impossible problem of deciding what level of time extension exactly suits the needs of every person.

To summarize: first, time extension apparently has significant effects on test performance; though it helps the disabled, it may also affect the performance of the non-disabled. Second, the diagnosis of reading disability, at least in Hebrew, is far from standard and objective, and thus cannot provide a reliable indicator of the need for accommodations. Lastly, the reading-disabled are the largest group among those who ask for test accommodations in a high stakes testing program. This state of affairs led us, at NITE, to try to accommodate the reading disabled with the use of a computerized test, while continuing to test the non-disabled in a P&P version of the same test. But this solution raises the question whether the computerized test is equivalent to the P&P version of the test. In the next section, we are going to answer this question, based on the results of a controlled experiment.

Effects of Testing mode, time constraints, and adaptivity on performance in computer based testing

The transition from a pencil and paper (P&P) test to a computerized adaptive (CAT) version of that test involves several factors, which affect examinees' behavior and the nature of the test.

In most cases, a location and scale difference between performance on a P&P version and a CAT version of the same test are not of interest as long as both versions measure the same construct at comparable levels of precision yet are used for separate purposes. However, when the two versions of the test coexist and are used interchangeably, one has to ensure that they measure performance on the same scale.

Of the many factors that might contribute to the difference in performance between a CAT and a P&P version of the same test, the three most salient seem to be:

1. Mode of testing: paper vs. computer;
2. The time constraints imposed on the examinee;
3. The adaptivity of the test or its lack thereof (adaptive vs. linear administration).

The purpose of the following experiment was to investigate the unique and combined effects of these three factors.

The subjects and the experimental design. Two hundred and sixty-seven subjects were tested in an experimental administration of four different versions of the Psychometric Entrance Test (PET), which until recently was administered regularly to all university applicants in Israel. All examinees were university applicants who were registered to take the PET for the first time about one month after the experimental administration. A random sample of 656 applicants was invited to participate in the experimental administration of PET. The letter of invitation emphasized the fact that this administration provided an opportunity to practice on questions similar to those included in the operational test (PET) they were registered to take. Of the 656 applicants, 267 agreed to participate. Six out of the 267 did not take the PET as

scheduled and were therefore excluded from most analyses, leaving a total sample of 261 examinees.

The Psychometric Entrance Test (PET) – is designed to assess abilities in three domains: verbal reasoning (V), quantitative reasoning (Q) and proficiency in English as a foreign language (E). The P&P operational version of PET consists of six tests, two per domain, each test containing 22-30 items that must be answered within 25 minutes.

The number right score in each domain is scaled to range from 50 to 150, with a mean of 100 and a standard deviation of 20. A total (TOT) score is computed by a weighted sum of the domain scores and scales back to a mean of 500 and standard deviation of 100. The relative weights of the three domains are 2, 2, and 1 for V, Q and E respectively.

Four experimental versions of PET, each including three tests one per each domain, were administered:

P&P – Paper and pencil linear administration with 25 minutes time constraint per test.

CTT – Computerized linear administration with 30 minutes per test.

CTI – Computerized linear administration with time limit per item.

CAT – Computerized adaptive administration with time limit per item.

All computerized versions of the experimental PET were administered using the MicroCAT software system (ASC, 1987). The CAT was based on a 3P-logistic IRT model, and a Bayesian prior likelihood function was used to estimate ability.

Two parallel forms (A and B) were used in the P&P, CTT and CTI experimental versions. In the CAT condition, a minimal and maximal number of items to be presented were predetermined for each section of the test. The test was terminated as soon as the error of estimation reached below a critical value, if the minimal number of items was presented; otherwise, the test was terminated with the presentation of the maximal number of items. The minimal and maximal number of items for each test was determined on the basis of results obtained in a previous

study (Ben-Simon, Sheffer, Ronnen and Cohen 1993), and ensured that most examinees would reach the critical value of the ability estimation error as predetermined for this particular test. The minimal and maximal numbers of items to be presented in each test were as follows: 30-36 items in the V test, 25-30 items in the Q test, and 25-30 items in the E test.

In the CTI and the CAT conditions, time constraints differed for items of different types, and were identical for all items of the same type. Time allotted for the various item types was determined on the basis of results obtained from previous administrations of computerized PET, and tended to be generous compared to the average amount of time given in PET and correspondingly in the P&P and CTT conditions.

The examinees were randomly assigned to one of the four experimental testing conditions P&P, CTT, CTI and CAT. Note that in each pair of consecutive conditions one differs from the other by one factor only. Therefore, the unique effect of each factor can be examined by comparing two consecutive testing conditions at a time, as follows:

Experimental testing Versions compared	Conditions compared	Effect examined
P&P vs. CTT	Paper & pencil vs. computer	Mode of testing
CTT vs. CTI	Per test vs. per item	Time constraints
CTI vs. CAT	Linear vs. adaptive	Item presentation model

In order to examine the unique effect of the mode of testing, the CTT condition should have been administered with a time limit of 25 minutes. Yet, in order to establish the exact time required for each test section administered under the CTT condition to match the results obtained under the P&P condition, an extra five minutes were added to each section in the CTT version. This generous time allocation, along with the precise recording of response latencies, permitted a separate analysis of the performance under different time constraints ranging from 25 to 30 minutes (hence, CTT25, CTT26… CTT30).

Results

Each examinee had four scores (V, Q, E, and a combined score – TOT) on the experimental test, as well as the scores on the operational PET test that was taken about one month after the experiment was over. Means and SD's of these scores for each condition of the experimental administration are presented in Table 2. In spite of the fact that examinees were randomly assigned to the various experimental test conditions, moderate, though not significant, differences were found between the average ability levels (operational PET scores) of examinees in the research groups. In order to control these differences, the operational PET scores were used as covariates in all further analyses that involve comparisons between means.

A separate analysis was carried out to test for the unique effect of the following three factors: mode of testing, time constraints and items presentation model. The effect of each factor was studied by comparing the mean score difference of the appropriate groups. Thus, the unique effect of mode of testing was examined by covariance analysis of the mean scores obtained under the P&P and CTT conditions, with the operational PET scores serving as covariates. In order to match all aspects of testing conditions excluding the mode of testing, mean scores on the P&P condition were compared with the CTT25 scores. No significant mode effects were found for the V and Q scores and for the combined (TOT) score. Yet, a significant mode effect ($F_{(2,129)}$= 7.0) was found for the E test, with CTT25 scores being markedly lower than the P&P scores.

In order to establish the exact time required for each test administered under the CTT condition to match the results obtained under the P&P condition, six scores were calculated for each examinee in each test corresponding to the 25-30 minutes time limit. The mean of each of these scores (CTT25 to CTT30) was then compared to the mean score obtained under the P&P condition. A separate analysis of covariance was carried out to test the significance of the differences between each of the above pairs of scores. Table 3 summarizes the results of this analysis. The three test sections differ with respect to the effect that additional testing time (0-5 minutes) has on the deviation of the CTT scores from the P&P scores. None of the six scores

(CTT25 to CTT30) calculated for the V section differed significantly from the P&P score. Yet, an additional two minutes' testing time was required to bring the CTT scores as close as possible to the P&P scores, thus reducing mode effect to nearly zero. No significant differences were found in the Q section between the CTT and P&P scores when 0-2 minutes were added to testing time in the CTT condition. Any additional time beyond that yielded significantly higher scores in the CTT condition. The closest scores for P&P and CTT were obtained under a 25 minute time limit. Thus, no additional time was required in the Q test to compensate for mode effect. A significant mode effect was found in the E section, namely, when the CTT25 scores were compared to the P&P score. Though all other scores calculated for the CTT condition (CTT26 to CTT30) did not differ significantly from the P&P score, an additional five minutes' testing time was required in order to reduce the mode effect in the E section.

To sum up, if a computerized (non-adaptive) Psychometric Entrance Test is to yield the exact scores as its P&P version, two extra minutes should be added to the Verbal Reasoning sections and five minutes to the English sections.

It might be argued that results obtained from a retrospective analysis of responses - calculating the correct number of responses given up to a certain time limit unknown to the examinees in advance - does not necessarily reflect the results that would have been obtained if examinees had been informed in advance about these limits. In other words, tighter time constraints could possibly have hastened the response pace and thus improved the performance. Still, one could counter-argue that examinees cannot possibly control the speed of their performance with such accuracy as to precisely meet a given time limit, and in most tests they tend to proceed as quickly as possible and use any remaining time for reviewing their responses. This argument is supported in part by the following findings observed for the CTT condition: 7%-14% of the examinees (depending on the test section) finished working on the test in less than 25 minutes; 35%-60% of the examinees had reached the last question of the test in less than 25 minutes and used some or all of the time left to review their responses, and 10%-30% of the examinees did not reach the last question even after 30 minutes.

The effect of time constraints and of adaptivity

The unique effect of time constraints was examined by comparing the mean scores obtained under the CTT and CTI conditions. In order to control the additional time required to compensate for mode effect, the following scores were used for the CTT condition: CTT27, CTT25, and CTT30 for the V, Q and E tests correspondingly. No significant time effect was discovered for the V and E tests. Yet, a significant time effect was found for the Q test ($F_{(2,129)}$= 10.1, p<.01) and for the total score – TOT ($F_{(2,129)}$= 12.3, p<.01). In both cases CTI scores were markedly higher than the CTT scores.

The unique effect of item presentation model (adaptivity) was examined by comparing the mean scores obtained under the CTI and CAT conditions. No significant effect for the adaptivity factor was found for the three test sections and for the total score.

Table 4 summarizes the results pertaining to the separate effects of mode of testing, time constraints and adaptivity. Also summarized in this table are the results for a combined effect of all three factors. The combined effect was examined by covariance analysis of the mean scores obtained under the P&P and CAT conditions. This is the most relevant effect if a parallel CAT version is to be developed to a given P&P test. As expected, significant effects were found for the Q and E tests and for the total score (TOT), with all scores on the CAT condition being considerably higher than those on the P&P condition. In light of the analysis of the unique effect of each of the three factors embedded in the combined factor, it is rather easy to track down the source of the observed effects: generous time limits in the CAT version contributed to significantly higher scores on the Q test and compensated only too well for mode effect in the E test.

Concurrent validity of P&P and computer based tests

Concurrent validity was estimated in all the experimental groups, using Pearson correlation coefficients, the criterion being the parallel scores achieved in the operational administration of PET. Table 5 summarizes these results for the three

test sections (V, Q and E) and for the total score (TOT). Nearly identical validity coefficients were obtained for the total score in all experimental conditions, ranging from r=.88 (for the CTI condition) to r=.92 (for the P&P condition). Similar results were obtained for the E section. Larger differences between validity estimates of the different test versions were found in the V and Q tests. The highest validity of the V section was found for the P&P and CAT versions, while the highest validity of the Q section was found for the CTT and CAT versions. Note that the CAT versions yielded consistently higher validity estimates across all test sections as well as the total score. What is referred to here as concurrent validity coefficient, in a different context may be referred to as test-retest reliability. In a prior study (Ben-Simon & Cohen, 1988) the test-retest reliability of the PET was estimated to be 0.90, based on the data of thousands of examinee who, after 14 months (on average) re-took PET. The results of a test-retest correlation of 0.92 (in the P&P group) found in the present study, gives credence to results of the experiment. The concurrent validity of 0.90 in the CAT condition supports the claim that the P&P and the CAT versions of the test measure quite similar concepts.

Development of a CAT that is equivalent to a P&P version

The experiment proved two points. First, extending the time limit of a test would lead to higher scores among the non-disabled. Hence, the practice of extending the time limit of a high stakes test may jeopardize the fairness of the test. Second, an adaptive, computer based version of the PET apparently measures the same construct as the P&P version while providing very generous time per item. But it also assigns higher scores than the P&P version. A suitable CAT version would then be one that is built on the P&P item pool, but with a score scale that, by proper equating process, is equivalent to the P&P score scale. This test would have the advantages of an adaptive (short and accurate) test, it would give ample time per item for the benefit of the reading disabled, and, at the same time, would not give unfair benefit to the non-disabled.

The development of an adaptive test that is based on Item Response Theory (IRT) requires first, the demonstration that the item pools are essentially unidimensional. Then, the item parameters are estimated according to the IRT model that is chosen

as a basis for the adaptive test. The PET, as already mentioned, comprises three sections: verbal reasoning (V), Quantitative Reasoning (Q) and proficiency in English as a foreign language (E). Item pools for these three areas were created by combining items from dozens of operational PET forms. The unidimensionality of the item pools was determined by various methods (Kaplan-Sheffer, Ben-Simon & Cohen, 1992; Tractinsky, Ben-Simon & Cohen, 1989; Ben-Simon, Tractinsky & Cohen, 1989). It turns out that the Q and E item pools are essentially unidimensional. In the V item-pool additional factors can be identified, but the eigenvalue of the first factor is so large relative to the second or third, that for practical purposes the item-pool can be considered unidimensional. The three parameters logistic model was adopted for the development of the adaptive test. A program, NITEST (Cohen & Bodner, 1989), was developed to estimate item parameters, together with a software package, NITECAT (Cohen, Bodner & Ronnen, 1989), for studying and simulating IRT models.

The Computerized PET (CPET) for the use of examinees with learning disabilities was developed on a software platform – NITESTER – that was developed at NITE. A special feature of this platform is that it lets the developer replace the actual examinee by a simulated examinee (a “simulee”). In this way, without writing a special program, the researcher can study the expected characteristics of the computerized test by running a simulation of hundreds of simulees that are sampled from a chosen ability distribution. Since not all aspects of behavior can be simulated, the researcher cannot gain information about the temporal characteristics of the test, but statistics such as the expected length of the test, the average measurement error, and the amount of item-exposure, can be easily obtained.

Simulations of CPET on 1500 simulees who were sampled from a population in which the ability level is normally distributed, proved that the test is quite accurate at recovering the true ability level of the examinees (Moshinsky & Ronnen, 1998). More realistic test of CPET and its comparability to the P&P version was provided by two experiments (Heller & Moshinsky, 1999; Moshinsky, 2000). The subjects in these two experiments were university applicants who were registered to take the PET about four weeks following the experiment. About half of them were assigned to take the CPET and the other half were assigned to take a P&P version. The results of the

experiments provided enough data to calibrate the CPET score scale with the P&P score scale (Heller & Ronnen, 2000). By correlating the scores in the experiment with the scores that were achieved one month later in the operational PET, it was possible to substantiate the claim that the CPET measures a construct similar to the construct measured by the P&P test.

Initial results from the application of CPET for examinees with disabilities

Administration of the CPET on a regular basis began in July 2000. About a month before the test took place, the examinees received a practice test on CD. This test was developed several years before (Ben-Simon et al., 1993), in order to provide prospective examinees with a means to get a fast estimate of their expected PET score, and proved quite useful in having the examinees familiarize themselves with the computer-based test.

Initial results of the operational administration were reported by Moshinsky & Kazin (2002) and pertain to 353 examinees who were entitled to the extended time limit, but were offered the opportunity to take the CPET. It is too early to get criterion validity data about the examinees, but it is possible to compare them, although with one reservation, with the non-disabled examinees who participated in the experiments. The reservation is that the two groups also differ with respect to the motivation in taking the test. While one group took the test as an operational, high stakes test, the other sat for it in a context of an (optional) experiment.

The mean scores of the examinees are displayed in Table 6. The LD group displays a pattern that is apparently typical of the learning disabled. The mean verbal reasoning score (V) is at about the same level as the quantitative reasoning score (Q), and both are significantly higher than the English proficiency score (E). This is in contrast with the flat score profile of the regular examinees. The mean total testing time (in minutes) for the two groups and the division of time among the test sections is also displayed in Table 6. It is interesting to note that the LD group took as much as 50% more time to answer the test. Although the two groups are not matched for ability, this might indeed prove that the problem of insufficient testing time for the LD was circumvented, as was intended.

Summary

We started by posing a general question – how to ascertain that test accommodations do not hamper equity and fairness. We suggested that providing a computer based test, and more specifically, an adaptive test, might circumvent the problems associated with extending time limits in a high stakes test. We discussed some of the issues in developing computerized adaptive tests and described the various stages in the actual development of a CAT version. The ultimate proof of the viability of the idea that we have explored will be data relating to the predictive validity of the CAT as compared with the validity of the P&P test. At the present we find that the proposed solution is satisfactory. The use of CAT for examinees with learning disabilities has facilitated the process of deciding whether to grant accommodations and has greatly reduced the cost of making wrong decisions.

Computer-based, or web-based, testing has many capabilities that are only beginning to be realized. Some of these capabilities might well be utilized in helping people with disabilities to compensate for and even circumvent their disabilities. One example is the capability having the computer “read aloud” selected sections of the test. We are now at the beginning of a research project that will test this option for the benefit of individuals with deep dyslexia. As in the case of time limit extension, the first question is whether the option of reading aloud might help (or maybe retard) the performance of regular examinees. We hope to be able to report initial results next year.

References

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

Birnbaum, A. (1968) Some Latent trait Models and their use in inferring an examinee's ability. In Lord, F.M. & Novick, M.R. (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Ben-Simon, A. (2001). *Learning Disabilities: A Psychometric Perspective.* A paper presented at the 27th annual IAEA conference, Rio de Janeiro.

Ben-Simon, A. & Cohen, Y. (1988). The effect of repeated testing on scores in Psychometric tests. (Hebrew) In: Nevo, B. & Cohen, Y. (Eds.) *Problems and Issues in Assessment and Measurement*. Jerusalem, Israel: NITE.

Ben-Simon, A., Ronnen, T., Sheffer, L. & Cohen, Y. (1993). *A CAT version of PET intended for initial estimation of ability.* Research Report No. 175 (Hebrew), Jerusalem, Israel: NITE.

Ben-Simon, A., Tractinsky, N. & Cohen, Y. (1989). *Item Banking of EFL items using the 3P logistic models.* Research Report no. 103. Jerusalem, Israel: NITE.

Cohen, Y., Ben-Simon, A. & Tractinsky, N. (1989). *Computerized adaptive test of English proficiency*. Research Report No. 98. Jerusalem, Israel: NITE.

Cohen, Y. & Bodner, G. (1989). *A Manual for NITEST – a program for estimating IRT parameters.* Research Report no. 94. Jerusalem, Israel: NITE.

Cohen, Y., Bodner, G. & Ronnen, T. (1989). *A manual of NITECAT – software package for research on CAT/IRT version 1.* Research Report no. 100. Jerusalem, Israel: NITE.

Eitan, M. (2002). *Operating principles in determining test accommodations*. Internal unpublished manuscript (Hebrew), NITE: Jerusalem.

Eitan, M., Moshinsky, A. & Ben-Simon, A. (In preparation), *A review of 200 LD diagnostic reports.* NITE: Jerusalem.

Hambleton, R.K. & Swaminathan, H. (1985). *Item Response Theory: Principles and applications.* Dordrecht, The Netherlands: Kluwer.

Heller, D. & Moshinsky, A. (1998). *CPET for examinees with disabilities – version1: results of experiment 1.* Research Report no. 256. (Hebrew), Jerusalem, Israel: NITE.

Heller, D. & Ronnen, T. (2000). *Calibrating the scores of the CPET for examinees with disabilities to the score scale of the P&P PET.* Technical Report no. 109 (Hebrew). Jerusalem, Israel: NITE.

Kaplan-Sheffer, L., Ben-Simon, A. & Cohen, Y. (1992*). A study of the dimensionality of the verbal reasoning item-pool.* Research Report no. 165 (Hebrew). Jerusalem, Israel: NITE.

Lord, F.M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, New Jersey: Erlbaum.

Moshinsky, A. (2000). *CPET for examinees with disabilities – version2: results of experiment 2.* Research Report no. 276. (Hebrew), Jerusalem, Israel: NITE.

Moshinsky, A. & Kazin, C. (2002). *Constructing a Computerized Psychometric Adaptive Test for University Applicants with Disabilities*. Paper presented at the AERA Annual Meeting. New Orleans, USA.

Moshinsky, A. & Ronnen, T. (1998). *CPET for examinees with disabilities – Test structure and simulation results.* Technical report no. 85 (Hebrew). Jerusalem, Israel: NITE.

Runyan, M.K. & Smith, J. (1991). Identifying and accommodating learning disabled law school students. *Journal of Legal Education, 41*, 317-349.

Tractinsky, N., Ben-Simon, A., Cohen, Y. (1989). *Goodness of fit to the 3P logistic model of the Quantitative Reasoning item pool.* Research Report no. 90 (Hebrew). *Jerusalem, Israel: NITE.*

Zuriff, G. E. (2001). Extra Examination Time for students with learning disabilities: An examination of the maximal potential thesis. *Applied Measurement in Education. 13*(1) 99-117.

Table 1:

Types of test accommodations provided by NITE to university candidates.

Reported disability	Accommodation
Visually impaired	• Recorded test • Special illumination conditions • Enlarged print • Closed circuit TV (magnifier) • Deletion of some item types
Auditory problems	• Instructions given by specially trained proctor
ADD/ADHD	• Testing in a separate room (fewer examinees) • Special breaks between test sections
Physical problems	• Special seating arrangements • Special breaks as needed • Help filling in the answer sheet
LD: reading	• Extended time • Recorded test
LD: mathematics	• Electronic calculator • Extended time
LD: writing	• Help filling in the answer sheet • Extended time
LD: understanding	• Not applicable

Table 2:

Means and SD's of the experimental PET scores, and of the experimental PET scores controlled for ability (operational PET scores).

Test section		P&P	CTT 30 min	CTI	CAT
Experimental PET scores					
V	mean	110	112	116	113
	std	18	21	15	14
Q	mean	108	110	116	109
	std	21	21	17	17
E	mean	109	105	114	109
	std	20	22	21	19
TOT	mean	553	556	588	560
	std	93	95	76	75
Operational PET scores					
V	mean	113	112	116	112
	std	17	16	17	18
Q	mean	114	111	113	109
	std	17	16	17	17
E	mean	111	108	114	107
	std	23	23	23	21
TOT	mean	573	563	581	557
	std	89	84	83	87
N		64	68	64	65

Table 3:

Means of the experimental PET scores (SSC) and of the experimental PET scores controlled for ability (C-SSC) as obtained under the P&P condition and under different time constraints (25-30 minutes) in the CTT condition.

Score		P&P	Time constraint (minutes)					
			25	26	27	28	29	30
V	SSC	110.9	108.3	109.4	110.6	111.1	111.7	111.7
	C-SSC	110.0	108.4	109.5	110.6	111.2	111.8	111.8
	F(2,129)		1.7	0.5	0.0	0.0	0.3	0.3
Q	SSC	107.7	106.8	108.2	108.9	109.6	110.3	110.9
	C-SSC	106.3	108.0	109.5	110.2	110.9	111.7	112.3
	F(2,129)		0.6	2.0	3.0	4.2*	6.0*	7.5*
E	SSC	108.7	101.9	103.1	104.3	105.3	105.8	106.2
	C-SSC	107.5	103.0	104.3	105.5	106.5	107.0	107.4
	F(2,129)		7.0*	3.4*	1.3	0.3	0.1	0.0
TOT	SSC	552.7	536.9	554.1	549.6	553.4	557.1	559.0
	C-SSC	547.7	541.6	548.8	554.4	558.2	562/0	563.9
	F(2,129)		0.7	0.0	0.9	2.2	4.1*	5.4*

* Statistically significant effect ($p<.05$)

Table 4:

Mode effect, time effect and adaptivity effect as found in an analysis of variance test applied to each analysis stage.

	V	Q	E	TOT
MODE EFFECT				
P&P vs. CTT25				
P&P	110.8	106.3	107.5	547.7
CTT25	108.4	108.0	103.0	541.6
F(2,129)	1.7	.6	7.0*	.7
TIME EFFECT				
CTT-EQ vs. CTI				
CTT-EQ	111.8	107.5	108.7	555.1
CTI	115.0	114.7	111.6	580/2
F(2,129)	2.6	10.1**	3.3	12.3**
ADAPTIVITY EFFECT				
CTI vs. CAT				
CTI	115.2	114.0	111.4	579.2
CAT	113.7	110.3	112.3	569.3
F(2,126)	1.0	3.2	.3	2.5
COMBINED EFFECT				
P&P vs. CAT				
P&P	110.8	105.5	107.2	545.9
CAT	112.6	110.9	111.0	566.4
F(2,126)	1.5	6.8*	6.5*	10.7**

Note:

- CTT25 indicates the CTT condition with 25 minute time constraints
- CTT-EQ indicates the CTT condition with different time constraints applied to each test in order to allow for equivalent scores on CTT and P&P
- * p<.05, ** p<.01

Table 5:

Pearson correlation coefficients between scores obtained on the experimental and operational administrations of PET (all the coefficients are statistically significant beyond the .05 level)

Test	P&P	CTT 30 min	CTI	CAT
V	.87	.75	.80	.85
Q	.75	.81	.62	.85
E	.90	.89	.92	.91
TOT	.92	.89	.88	.90
N	64	68	64	65

Table 6:

Mean scores and test-duration of regular examinees and examinees with disabilities, for three sections of the test and the total test.

	Regular examinees (n=338)	Examinees with disabilities (n=353)
V mean score	110	108
Q mean score	112	110
E mean score	112	103
Mean total score	565	544

Mean duration of the test (in minutes)	137	198
% time devoted to V section	31%	33%
% time devoted to Q section	39%	35%
% time devoted to E section	22%	22%
% time devoted to instructions and breaks	8%	10%