דוח המרכז

361

Applications of CAT in Admissions to Higher Education in Israel: Twenty-Two Years of Experience

Naomii Gafni
Yoav Cohen
Keren Roded
Michal Baumer
Avital Moshinsky

June 2009

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 361

ISBN:978-965-502-152-3

Graduate Management Admission Council®

Applications of CAT in Admissions to Higher Education in Israel: Twenty-Two Years of Experience

Naomi Gafni, Yoav Cohen, Keren Roded, Michal Baumer, and Avital Moshinsky

National Institute for Testing and Evaluation
Jerusalem, Israel

Presented at the CAT in Spain and Israel Paper Session, June 2, 2009

Abstract

The use of CAT in higher education admissions testing in Israel is described. This includes: (1) AMIRAM—a CAT of English as a foreign language that has been used by various institutions of higher education for placement purposes for the past 22 years, and (2) MIFAM—a CAT version of the Psychometric Entrance Test that has been in use for nine years as a higher education admissions tool for examinees with disabilities. Both applications run in parallel with paper-and-pencil test versions. This presentation focuses on the specific procedures used to produce equitable scores across the two media as well as examining the suitability of CAT for examinees with disabilities. Also discussed are a number of practical issues that were encountered during conversion of the Psychometric Entrance Test (PET) to a CAT format. Issues that pertain to the meeting of content specifications, item exposure, item banks, item bank dimensionality, and equating, are identified and discussed in the context of evolutionary changes in the MIFAM program.

Acknowledgment

Presentation of this paper at the 2009 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Gafni, N., Cohen, Y., Roded, K., Baumer, M., & Moshinsky, A. (2009). Applications of CAT in admissions to higher education in Israel: Twenty-two years of experience. In D. J. Weiss (Ed.), *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Naomi Gafni, National Institute for Testing and Evaluation, P.O.B. 26015, Jerusalem, 91260 Israel. Email: Naomi@nite.org.il

Applications of CAT in Admissions to Higher Education in Israel: Twenty-Two Years of Experience

The National Institute for Testing and Evaluation (NITE) is a non-profit organization responsible for test development, test administration, test scoring, and reporting of scores to both examinees and the various institutions of higher education in Israel. NITE was established in 1982 by the Israeli Universities Board. The first version of the Psychometric Entrance Test (PET) was administered by paper-and-pencil (P&P) in 1983. The following year a decision was made to develop a CAT based on IRT and the Unit for Computerized Tests was established. At the time it was quite clear that CAT was not going to be used extensively. The rationale for establishing the unit was that computerized testing would surely be used in the future, hence warranting the investment. The first operational CAT was administered in 1987. NITE serves as its own vendor.

This paper focuses on adaptive test applications, though NITE does administer other computerized non-adaptive tests: (1) MEMAD—an Internet-based admissions and placement test taken by candidates for preparatory colleges (7,000 examinees per year); (2) An audio version of PET used to examine sight-impaired candidates (230 examinees per year); (3) MATAL— test used to diagnose learning disabilities and attention disorders (about 2,000 examinees thus far); (4) Internet-based practice tests for PET (28,000 visitors to the Web site last year, 4,440 of which completed the test) and MEMAD (5,040 visitors, 1,097 of which finished the test during the last nine months). Overall, about 22,000 people are examined by computer each year.

Overview of the Psychometric Entrance Test

The Psychometric Entrance Test (PET) is a scholastic aptitude P&P test constructed and administered by NITE. It is used in the admissions procedures of all Israeli universities and colleges, in conjunction with the matriculation certificate. PET is comprised of three multiple-choice domains: Verbal Reasoning—V (60 items), Quantitative Reasoning—Q (50 items), and English as a Foreign Language—E (54 items). PET is translated into Arabic, Russian, English, French, and Spanish. A combined Hebrew and English version is offered to applicants who are not proficient in any of the aforementioned languages.

The P&P operational version of PET consists of six sections, two per domain, each containing 25-30 items that must be answered within 25 minutes. The number-correct 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 scaled back to a mean of 500 and standard deviation of 100, with a range of 200-800. The relative weights of the three domains are 2, 2, and 1 for V, Q and E respectively. Approximately 80,000 examinees take the examination annually and PET scores are reported to about 60 institutions.

A validity study conducted by NITE and based on data from 168,881 students studying at Israeli universities during 1990-1999 (Kennet, Bronner, and Oren, 1999) indicated that:

1. In most fields of study, the predictive validity of PET was slightly higher than that of the matriculation (0.45 and 0.41 for PET and the matriculation, respectively).
2. The Composite Score, which is based on both PET and the matriculation certificate,

There are 5 main administrations of PET every year. In each administration between one and three forms are administered. The PET item pool contains about 15,000 items.

CAT at NITE

Decisions regarding the development of CAT in Israel have been affected by several considerations (pertaining to the P&P versions but also pertinent to the development of CAT):

1. Israel is small, both in terms of geographical area and population size. The number of examinees restricts the number of new test forms that can be pre-tested and constructed annually (a maximum of 12 forms). The number of new test forms affects the number of test administrations, the number of examinees per test form, and the number of test centers and computer stations. All of these affect the final cost.
2. The PET is a high-stakes admissions test used by all the universities in Israel. Coaching is popular in Israel and about 80% of the examinees participate in test preparatory courses. Therefore, the policy of NITE is to minimize reuse of previously administered test forms. Recently, the Israeli parliament issued a law according to which NITE is required to disclose one test form each administration date— five test forms annually. In order to implement CAT securely, there is a need to develop no less than 12 P&P test forms annually.
3. Hebrew test forms are translated into five languages. The PET is administered in Arabic four times a year. To produce one verbal reasoning test form in Arabic, about two test forms in Hebrew are necessary. This means that at least eight new test forms in Hebrew must be developed annually.
4. NITE's budget is based on a test registration fee approved by the board of directors every year. It is evident from the relevant literature (e.g., Rudner, 2007) and our own financial estimates, that in order to implement an efficient and secure computerized testing system, NITE requires a much larger cash reserve than it currently has at its disposal.

Hence, while NITE has continued to develop the infrastructure for CAT, the applications have been restricted to specific uses and populations. These include: (1) AMIRAM (the English as a foreign language CAT); and (2) MIFAM (the CAT version of the PET). Both applications run in parallel with paper-and-pencil test versions.

Overview of the MIFAM – An Adaptive Computerized Version of PET for Examinees with Disabilities

Over the past few years, there has been a large increase in the number of university applicants requesting special testing accommodations for university entrance exams (Camara, Coperland, & Rothschild, 1998; Moshinsky and Kazin, 2005). In 2008, about five percent of applicants requested testing accommodations on PET and 65% of them were granted accommodations. The increase has brought to the fore certain psychometric issues pertaining to the fairness of testing students with disabilities and the comparability of special and standard testing conditions. For example, providing accommodations on the P&P test, such as increased time, yields numerous nonstandard administrations.

To address these problems of fairness and standardization, as well as to accommodate the significant increase in the number of university applicants requesting accommodated testing (whether for learning or physical disabilities), NITE developed a psychometric CAT (MIFAM) with a generous allotment of time per item (rather than per section or test). The CAT both accommodates the special needs of a majority of candidates with disabilities and is comparable to a test administered under standard conditions. Although MIFAM is currently used mainly by students with disabilities, it nonetheless incorporates the principles of universal test design, addressing the needs of all examinees, both with and without disabilities (Cohen, Ben-Simon, Moshinsky & Eitan, 2002, Moshinsky & Kazin, 2005). Using MIFAM also removes the problem of flagging because there is no longer a group of students taking the same test under nonstandard conditions. It offers the ability to allocate more time to students while lightening the total time burden because the test requires fewer items than would a non-adaptive test with an additional time accommodation.

Prior to constructing MIFAM, NITE surveyed students and found no significant differences in attitudes toward computers between students with and without learning disabilities (Moshinsky, Tenebbaum, Rapp, & Ronen, 1997). The first operational administration of MIFAM took place in July 2000. To date, 5,100 candidates have been tested with MIFAM (about 1,000 of them in 2008). In feedback questionnaires, examinees with disabilities reported satisfaction with the CAT and rated it clear and user friendly. They also judged the test to be generous in time allotment, compensating for slow reading and difficulties in concentration; and they perceived the test to be fair.

MIFAM is installed on portable computer labs which are transferred on testing days to testing centers around Israel. MIFAM is a high-stakes test and hence prone to security breaches. There are three administrations of MIFAM for candidates with disabilities per year. Usually, two new MIFAM forms are administered per year.

Once the CAT became operational, the performance time, scores and quality of responses to different types of items of examinees without disabilities and examinees with disabilities could be compared. The data accumulated thus far suggest that the CAT enables examinees with disabilities to demonstrate their potential as well as, and even better than, the P&P test with accommodations. Analysis of the CAT data shows that performance time of examinees with learning disabilities and those with physical disabilities is significantly longer than the performance time of examinees without disabilities. This is consistent for all three domains of the test and for all item types.

Overview of the AMIRAM—An Adaptive Computerized Version of the E Subtest

The E subtest of PET serves a dual purpose: it is a component of the PET total score, and it is also used for placement of students in English as a foreign language classes. The AMIRAM is an adaptive computerized version of E used for placement purposes. It contains three types of questions: sentence completions, restatements, and reading comprehension. About 12,000 examinees take the examination annually. Around 117,000 examinees have taken the test so far. AMIRAM is a medium-stakes test and therefore less prone to security breaches.

Registration for AMIRAM is carried out by the various institutions. The AMIRAM is administered on site at the institutions, using portable computers. The computers are supplied by

NITE and accompanied by test coordinators and proctors. One of six versions of the AMIRAM is installed on each computer. The versions change periodically. The allocation of the different AMIRAM versions to the different computers and the allocation of the computers to the examinees are random. This procedure ensures security.

Implementation Issues

CAT at NITE relies on the 3-parameter item response theory model. In order to implement CAT at NITE, a software package (NITECATSYS) was developed. The package consists of a number of programs that allow test generation, quality assurance, and the administration of a CAT (Blum & Ronen, 2001). There are two additional modules that run with the NITECATSYS in administrations of the test: the Human-Machine Interface (HMI) module and the administration module (Zach & Rahamim, 1999). Items for the CAT are selected from the standard P&P test.

Unidimensionality

The first step in the process of implementing CAT was to investigate the unidimensionality of each of the three domains (Kaplan-Sheffer, Ben-Simon, & Cohen, 1992; Trackinsky, Ben-Simon, & Cohen, 1989; and Ben-Simon, Trackinsky, & Cohen, 1989). The procedure was based on the method developed by Rosenbaum (1984) and on factor analysis. It was determined that the three sections fulfill the requirements of the model.

Parameter Estimation

Parameter estimation is based on operational data of the P&P test. At first parameter estimation was conducted through using ASCAL (Assessment Systems Corporation, 1987), and then by NITEST—an estimation program that was developed by NITE (Cohen & Budner, 1989) in order to handle large numbers of items concurrently. Since 2002, parameters have been estimated by means of BILOG-MG (Zimowski, Muraki, Mislevy & Bock, 1996).

Test Structure

In the CAT, the first two items in each domain have an average level of difficulty and low discrimination. These items are randomly sampled, the first from a unit of six and the second from a unit of three items, in case of a correct response, and another unit of three items in case of an incorrect response. The rationale is that the first two questions should be neither too difficult nor too easy, and that the posterior variance should not decrease too much after the first two items. The test reaches its end once an examinee satisfies either of the following criteria: (1) the posterior variance is smaller than a predefined value, and the examinee has finished the minimum number of questions as defined or (2) the examinee has reached the maximum number of items. The test structure and the time allotted to each item type are detailed in Table 1. In the CAT, examinees cannot return to earlier items, change answers they have submitted, or skip items. They can, however, not respond during the allotted time.

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 use of different sampling rules for different units controls the exposure of items and thus provides good psychometric indices of the test. The adaptivity of the procedure holds within each item type group. Within an item set (e.g., five items related to the same reading comprehension paragraph/graph/table) there is no adaptivity

(The item set itself is selected based on maximum information/match to item difficulty).

Table 1. Computerized Adaptive PET Structure: Test Domains and Item Types, Percentage of Items of Each Item Type, and Time Allotment Per Item Type

Domain	Item Type	P&P	CAT	Time Allotment per Item (in Min.)
Verbal Reasoning	Words and phrases	~13	10-15	1.0
	Verbal analogies	~20	13-31	1.5
	Letter Switching	~13	12-16	3.0
	Sentence Completions	~17	17-23	3.0
	Logic	~17	17-23	4.0
	Reading Comprehension	~20	12-16	7.0 (per text), 4 (per item)
Total Computerized				100.0-118.0
Total P&P				50.0
Quantitative Reasoning	Questions and Problems	~60	57-66	4.0
	Diagrams & tables	~16	11-14	5.0 (per graph), 4.0 (per item)
	Quantitative Comparison	~24	23-29	4.0
Total Computerized				117.0-157.0
Total P&P				50.0
English	Sentence Completions	~41	38-54	2.0
	Restatements	~22	29-38	4.0
	Reading Comprehension	~37	18-24	7.0 (per text), 4.0 (per item)
Total Computerized				75.0-89.0
Total P&P				50.0

Prior to making the test operational, simulations are run in order to ensure that (1) item exposure is not too high for certain items and too low for other items and (2) that a high proportion of the examinees reach a predetermined minimum level of posterior variance.

The exposure rates are examined for each item type. If, for a certain item type, the exposure rate is too low, one of the possible solutions would be to increase the number of items selected from this item type. If, on the other hand, the exposure rate is too high, it might be possible to reduce the number of selected items or to add items of this type to the item pool. It is also possible to change the item selection rule; if items are available at all levels of difficulty, one

would choose the "match to b" selection rule. If the items have a similar difficulty level, they could be selected randomly or according to the combination rule.

For example, assume that the pool of a certain item type contains 10 items: item 1, item 2, item 3…item 10. Two items of this item type have to be selected. If the simulation indicates that item 10 is over-exposed, it is possible to divide the items into two groups, from each of which one item would be selected: items 1-9 and items 1-10. This procedure would decrease the exposure rate of item 10.

Posterior variance is examined as a function of ability (θ) level. If the simulation indicates a relatively high proportion of examinees of a certain θ level with a high level of posterior variance, items of the appropriate difficulty level are added to the pool. It is always possible to either increase the minimal number of items presented to the examinees or increase the number of items in the pool at certain levels of θ. Sometimes it is necessary to change the order in which the various item types are presented to the examinees. It should be noted that certain items are more vulnerable to exposure and easier to memorize than others and should therefore be treated with more caution.

Content Specifications

Since the P&P and CAT versions of PET are used simultaneously, it was decided to construct the CAT with the same content constraints as the P&P test. The pool is organized according to item type.

The number of CAT forms that are developed per year is limited by the number of operational P&P forms that are developed every year. Each CAT form is based on three or four P&P test forms, supplemented by additional items if necessary. Each item pool contains about 210 items for V, 175 items for Q, and 230 items for E. These figures are similar to those mentioned by Rudner (2007) for the item pool size needed in order to incorporate a realistic set of constraints and standard error targets when developing a CAT version of the GMAT. Simulations are run to ensure the quality of the new forms both in terms of convergence rates and item exposure rates.

Special Characteristics of the CAT

The most significant respect in which the NITE CAT differs from the NITE P&P test, as well as from other computerized tests, is that time is allotted per item and not per section or subtest. This feature permits standardization at the level of a single item. A generous allotment of time per item was empirically determined in several experiments according to the actual latency distribution at the level of the item type (Rapp, Ronen, & Cohen, 1996). In the CAT, examinees get about one-half the number of items and an extended time allotment of about 100% to 400% compared to the P&P test. Using CAT provides a number of meaningful advantages in a standardized “package deal:” additional time allotment per item, fewer items, and a user-friendly HMI that includes use of a large font, separate presentation of each item, and rest breaks. Table 1 shows the computerized adaptive PET structure: test domains and item types, percentage of items of each item type, and time allotment per item type (note that the actual response time might be much shorter than the time allotted).

Scaling the CAT to the P&P Test

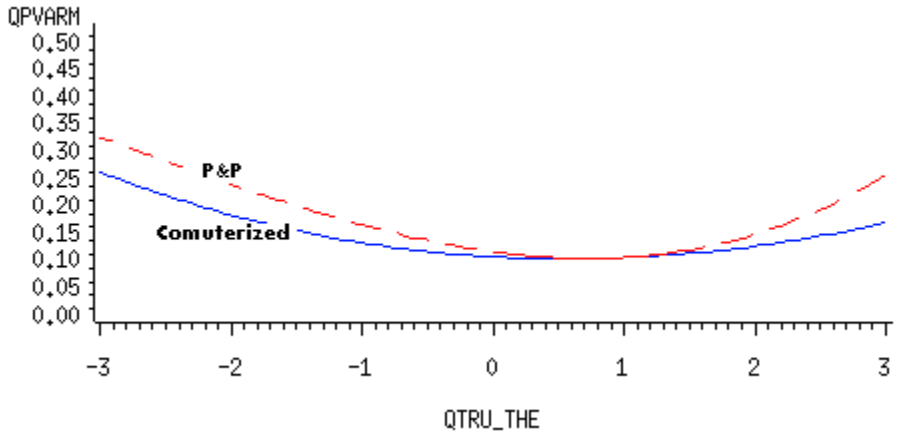
When P&P and CAT versions of a test coexist and are used interchangeably, one has to ensure that they measure performance on the same scale. The issue of extended time is the most significant source of noncomparability. Calibration between P&P scores and CAT scores is based on previous experiments (Ben Simon, Sheffer, Ronen, & Cohen, 1993; Cohen, Ben-Simon, Moshinsky and Eitan, 2002; Heller and Moshinsky, 1999; Moshinsky, 2000). The scaling of the scores from θ scores to comparable P&P scores was described by Moshinsky and Kazin (2005). The calibration was conducted in three stages: (1) final θ scores were transformed to a concordant number-correct score (the equations used for this procedure were the test characteristic curve and the transformation that equates the θ estimates on the CAT into number-correct scores on a P&P version of the test); (2) number-correct scores were transformed into standardized scores; and (3) additional corrections for the different modality and "doglegging" at the lower and the upper ends of the θ distribution were used.

Test simulations showed that the CAT had satisfactory psychometric properties when compared to the P&P test by means of posterior variance and the correlation between the true and estimated θ (Table 2 and Figure 1).

Table 2. Results of Test Simulations: Number of Questions Administered, Correlations of True and Estimated θ, and Posterior Variance (PV) and Its Standard Deviation (SD) for Three Intervals of θ

Modality and Test Area	Number of Questions		Correlation Between True and Estimated θ	$\theta < -1.5$ ($N = 97$)		$-1.5 < \theta < 1.5$ ($N = 1{,}304$)		$1.5 < \theta$ ($N = 97$)	
	Mean	SD		PV	SD	PV	SD	PV	SD
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
Computerized 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

Figure 1. Simulation Results (Quantitative Reasoning): A Comparison of the Posterior Variance as a Function of θ in the P&P Test and in the CAT



Three experiments were conducted to investigate the comparability of the CAT and the P&P test. In the experiments, a random sampling of applicants who had registered for the operational test were invited to participate in an experiment in return for an early estimation of their score on the operational test. The results of the experiments were very similar: they showed that scores on the CAT were similar to scores on both the experimental and operational P&P tests. The correlations between the scores on the experimental and the operational test showed that the CAT predicted operational scores as well as the P&P test did. The correlations were similar to the usual test-retest correlations found for the general population (Tables 3 and 4).

Table 3. Mean and SD of Scores for Computer and P&P Groups in Experimental and Subsequent Operational Tests

	Experimental Scores				Operational Scores			
	Computer Group (N = 338)		P&P Group (N = 329)		Computer Group (N = 338)		P&P Group (N = 329)	
Test	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Verbal Reasoning	111	18	113	20	116	19	116	19
Quant. Reasoning	113	9	111	17	118	18	118	17
English	113	22	114	22	116	21	117	21
Total	569	96	573	94	598	98	598	94

Table 4. Test-Retest Correlations for Experimental (P&P) and CAT Groups and for Operational Test Scores in the General Population

Test	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

Two additional variables were examined: gender and previous experience with computers. The results showed that the interaction between modality and gender was not significant, as was the interaction between modality and computer use (Moshinsky and Kazin, 2005).

The Development Stages of a CAT Version

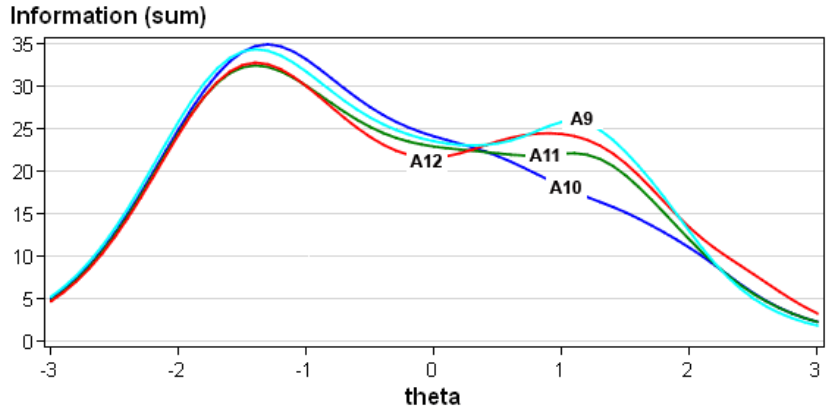
Each CAT version of MIFAM/AMIRAM is examined thoroughly before, during, and after administration. Lessons learned from these examinations are applied in developing the next version of the test. By way of example, the development of a recent version of AMIRAM is described below.

Quality checks during the development of a new version, prior to administration:

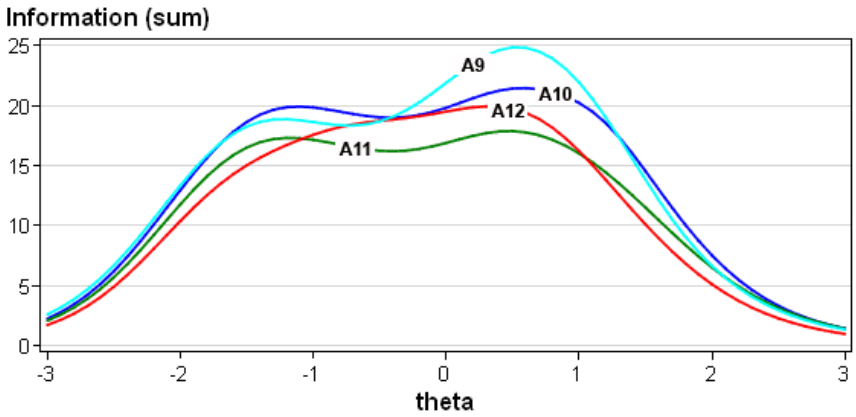
1. Select three P&P forms of E and additional source sections to complete the new version as necessary. Define enemy item sets, as well as other constraints.
2. Examine the information function of the test by item type and compare to previous versions (for example, see Figure 2).
3. Make sure that the item parameter values are correct and consistent in the system.
4. Examine the conversion table from θ to standardized scores (these tables are compared to previous tables).
5. Prepare installation software.
6. Determine the order of item presentation according to item type. For example, restatement items are usually presented before reading comprehension items. However, because reading comprehension items are sampled as item sets, the adaptive process is compromised by this order. This is of particular concern toward the end of the test when the final θ estimate is determined. Restatements, which are selected individually, are better facilitators of adaptivity. Moreover, the item pool contains fewer reading comprehension paragraphs to select from than restatement items.
7. Determine the rules for ending the test. Recently, it was decided to end the test for examinees whose θ estimate is beyond certain cutoff points (above or below), even if the posterior variance does not reach the predetermined value. In these cases, there is much confidence regarding placement decisions and the test ends after 23 items. Thus, examinees whose scores are at the extreme ends of the score distribution are not presented with additional redundant items that do not match their θ estimates, and unnecessary exposure of items is avoided.
8. Perform a final check of each item including key and instructions.

Figure 2. Information Functions for Four AMIRAM Versions (A9-A12) for Each Item Type and Across All Items

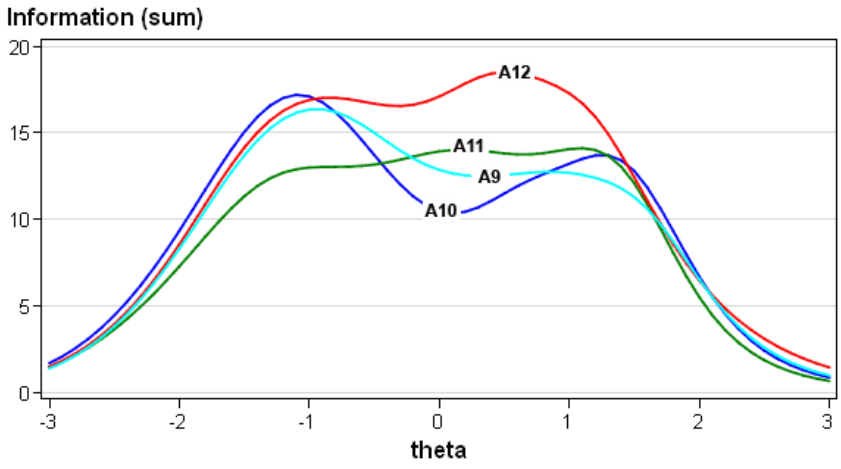
a. Sentence Completions



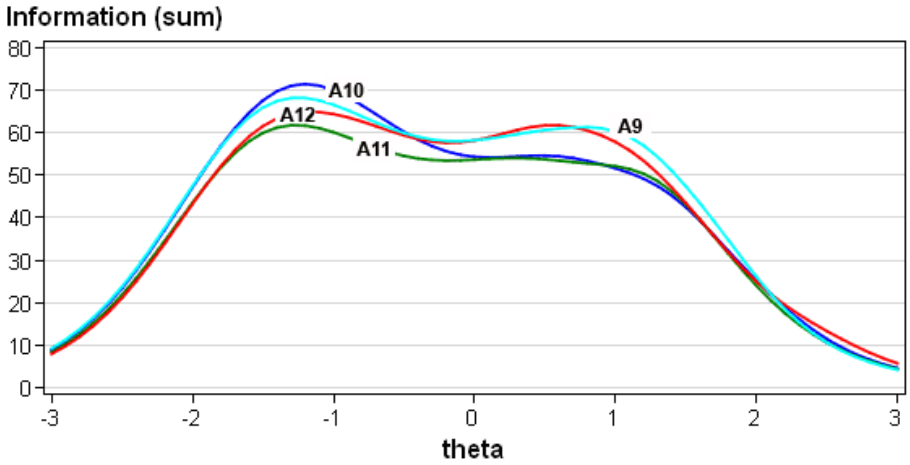
b. Reading Comprehension



c. Restatements



d. All Items



9. Identify possible problems in the testing administration and scoring: time allotment, instructions screens, scoring, and recovery process by running the test from the beginning to the end at least twice (using different people).
10. Run simulations. The following issues are examined by running simulations:
    a. Distribution of the converted standardized scores. This distribution is compared to previous distributions.
    b. Percentage of examinees for whom the posterior variance was smaller than 0.08 (expected value is 92%).
    c. Mean and SD of estimated θ, true θ, and posterior variance.
    d. Mean and range of test length (number of items.)
    e. The correlation between estimated θ and true θ, as well as between estimated θ and percent correct.
    f. The standard score when there is no mistake, two mistakes, or no correct answer.
    g. Make sure that the test ends according to the stopping rules.
    h. Exposure rates according to various criteria for each item.
11. Erase previous versions of the test from the portable computer disks.
12. Prepare a backup disk.
13. Document all changes, novelties, and decisions regarding the new version.

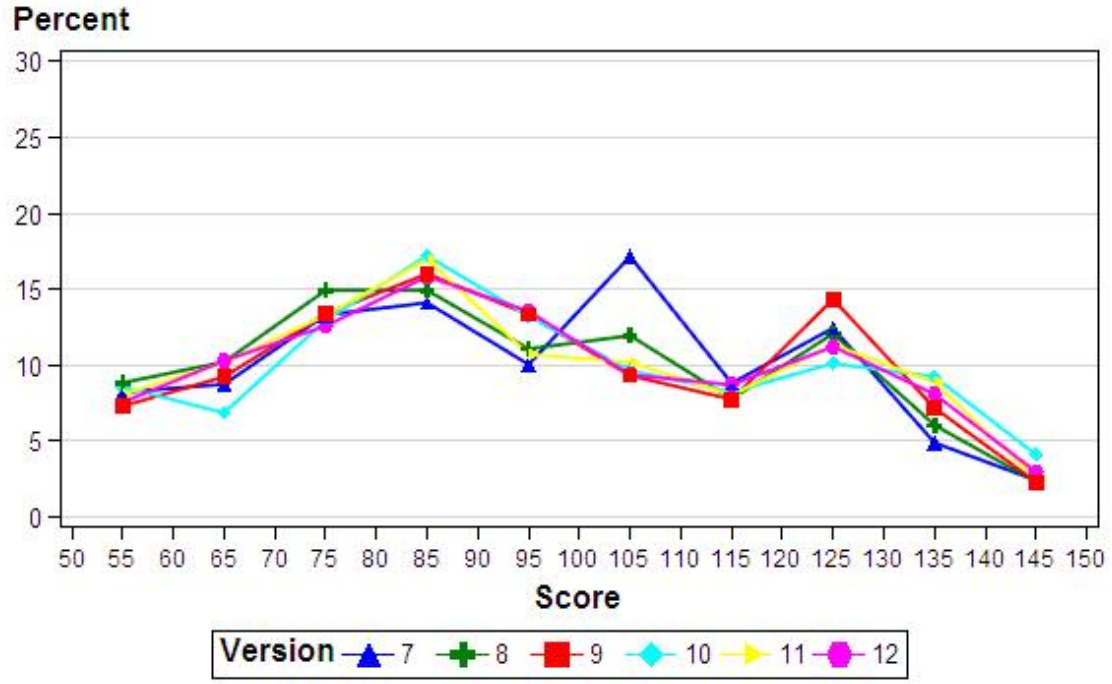
The data are examined once per predetermined period of test administration, so that problems can be identified and solved immediately. This examination process should be short, clear and fast.

Quality checks during the administration period of a CAT version:

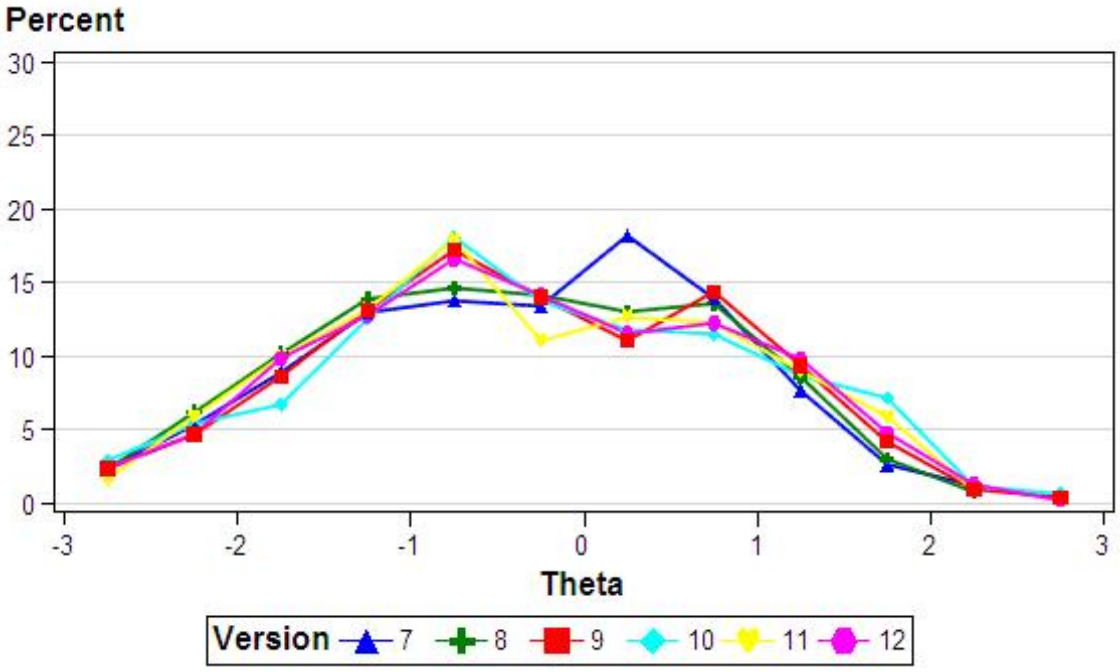
1. Extract data from the database.
2. Check the number of test administrations.
3. Check the score distribution by various variables (CAT version, college, computer lab). For example, Figure 3 shows the score distributions of six AMIRAM versions as a function of standardized scaled scores (Figure 3a) and as a function of estimated θ (Figure 3b).

Figure 3. Score Distribution by Version

a. Standardized Scaled Scores



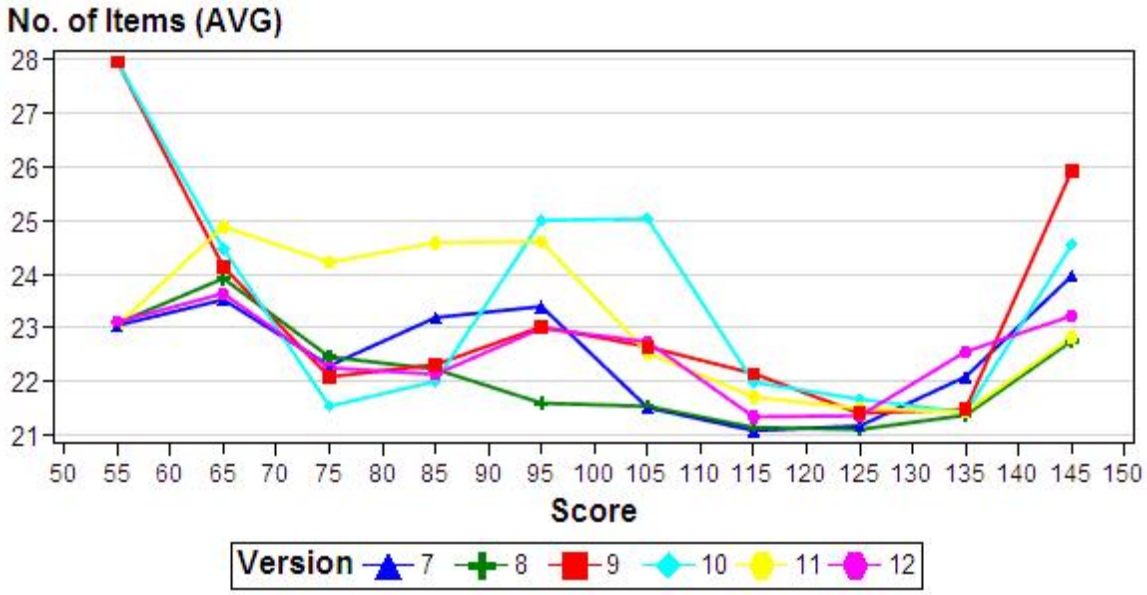
b. θ Estimates



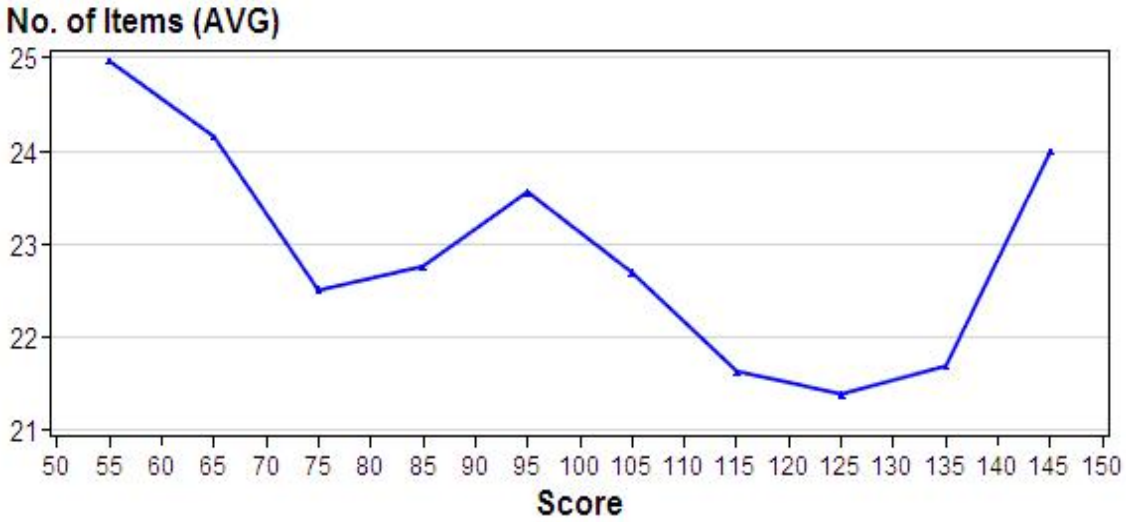
4. Examine θ values vs. number (percent) correct.
5. Examine number of incorrect answers vs. the standard score required for exempt level and for the maximum standard score of 150.
6. Check the number of items per examinee (minimum and maximum). Figure 4 shows the number of items per examinee for six AMIRAM versions.

Figure 4. Number of Items per Examinee by Version and Across Versions

a. The Number of Items per Examinee by Version



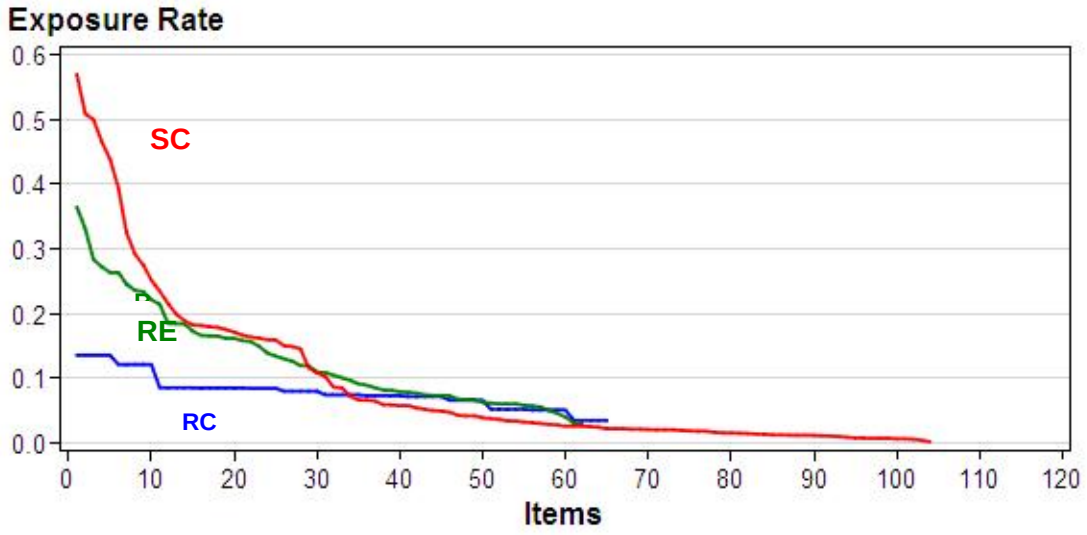
b. The Number of Items per Examinee Across Six Versions



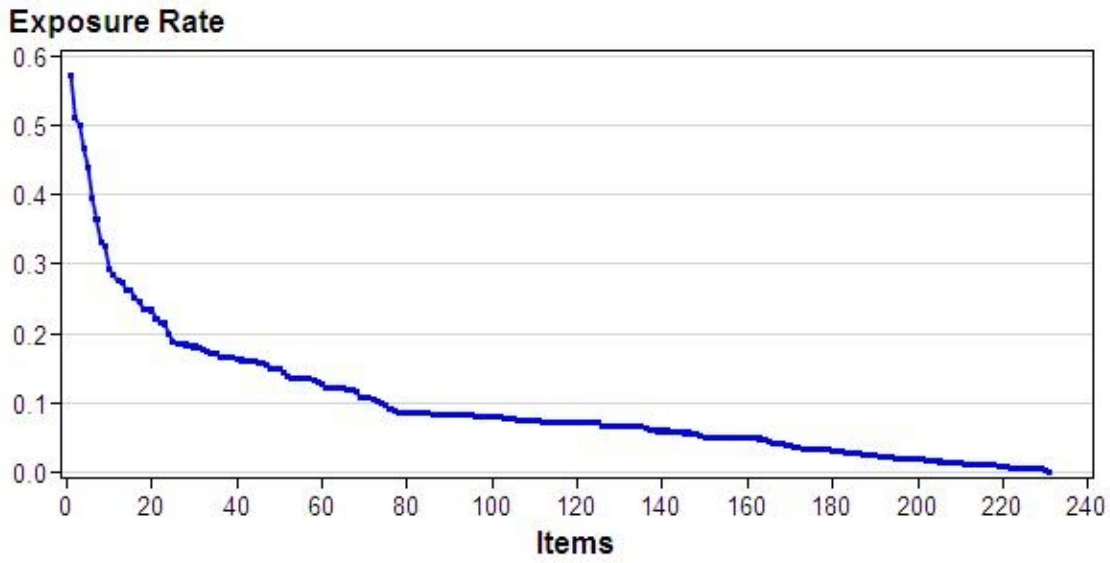
7. Check the percentage of examinees for whom the pre-determined posterior variance was or was not achieved.
8. The number of items administered for examinees according to their posterior variance.
9. Check for items that have not been exposed at all. Figure 5 shows exposure rates by item type (Figure 5a) and across item types (Figure 5b) for version 12.

Figure 5. Exposure Rates by Item Type for Version 12

a. Exposure Rates by Item Type for Sentence Completion (SC), Reading Comprehension (RC) and Restatements (RE)



b. Exposure Rates Across Item Types for Version 12



10. Examine correlation between exposure rates found in the simulation vs. practice.
11. Identify examinees with irregular high reaction times.

Conclusions

This paper presented some of the practical issues considered by the NITE in the design and evaluation of the CAT versions of the Psychometric Entrance Test (MIFAM) and of the English as a Foreign Language Placement Test (AMIRAM).

Some of the key considerations are:

1. *The need to minimize security risks* has limited the use of CAT to two applications: PET for examinees with disabilities and English for placement purposes. The first applies to relatively small populations and the second is not a high-stakes test.
2. *Both applications run in parallel with paper-and-pencil test (P&P) versions*. Therefore, content specifications should assure similarity across both versions and scores should be equivalent. Results obtained thus far have confirmed equivalence.
3. *Item pool characteristics*. Once again, due to security concerns, it was decided to employ several smaller item banks rather than one comprehensive bank. Each of these smaller banks is comprised of two to three P&P test versions, thus maintaining the same content specifications as the P&P test.
4. *The CAT algorithm*. In service of the dual objectives of maximal accuracy and controlled exposure of items, a variety of item selection parameters are employed. The final algorithm is determined on the basis of simulations.
5. *Time allotment and test length.* Unlike many other CAT systems, both applications feature per item time allotment and variable test length.
6. *Quality control.* Since the functioning of CAT is somewhat opaque, it is vital to articulate and follow stringent regulations thus ensuring that each examinee indeed receives the score that reflects his/her actual performance.
7. *Relevance.* Both applications have been operational for many years and have proven efficient and worthwhile. In particular, they provide suitable accommodations for examinees with special needs.
8. *Satisfaction.* Examinee feedback indicates high levels of satisfaction.

References

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

Ben Simon, A., Sheffer, L., Ronen, T., & Cohen, Y. (1993). *A computerized adaptive version of the PET for self evaluation of ability level.* (National Institute for Testing and Evaluation Report No. 175). Jerusalem, Israel: National Institute for Testing and Evaluation.

Ben-Simon, A., Trackinsky, N., & Cohen, Y. (1989). *Item-banking of EFL items using the 3-P logistic model* (National Institute for Testing and Evaluation Report No. 103). Jerusalem, Israel: National Institute for Testing and Evaluation.

Blum, N., & Ronen, T. (2001). *A technical review of the operation module in the NITECATSYS computerized test package* (National Institute for Testing and Evaluation Report). Jerusalem, Israel: National Institute for Testing and Evaluation.

Camara, W. J., Coperland, 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). New York: The College Board.

Cohen, Y., & Budner, G. (1989*). A manual for NITEST – A program for estimating IRT parameters* (National Institute for Testing and Evaluation Report No. 94). Jerusalem, Israel: National Institute for Testing and Evaluation.

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.

Heller, D., & Moshinsky, A. (1999). *Computerized adaptive test for the examinees with disabilities--Version 1: Results of Experiment 1* (National Institute for Testing and Evaluation Report No. 256). Jerusalem, Israel: National Institute for Testing and Evaluation.

Kaplan-Sheffer, L., Ben-Simon, A., & Cohen, Y. (1992). *Item-banking of verbal reasoning items using the 3-P logistic model* (National Institute for Testing and Evaluation Report No. 165). Jerusalem, Israel: National Institute for Testing and Evaluation.

Kennet-Cohen,. T., Bronner, S. & Oren, C. (1999). *A meta analysis of the predictive validity of the selection process to universities in Israel*. Megamot, 40, 54-71 (In Hebrew).

Moshinsky, A., & Kazin, C. (2005). Constructing a computerized adaptive test for university applicants with disabilities. *Applied Measurement in Education, 18,* 381-405.

Moshinsky A., Tenebbaum, M., Rapp, Y., & Ronen, T. (1997). *The habits of using computers among examinees with learning disabilities and physically handicapped* (National Institute for Testing and Evaluation Tech. Report No. 72). Jerusalem, Israel: National Institute for Testing and Evaluation.

Rapp Y., Ronen, T., & Cohen, Y. (1996). *Analyzing item type performance time in an experimental administration of the CPETSE* (National Institute for Testing and Evaluation Tech. Report No. 52a).). Jerusalem, Israel: National Institute for Testing and Evaluation.

Rudner, L., M. (2007*). Implementing the Graduate Management Admission Test® Computerized Adaptive Test.* In D. J. Weiss (Ed.), Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing. Retrieved from www.psych.umn.edu/psylabs/CATCentral/ .

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, *49*, 425-435.

Trackinsky, N., Ben-Simon, A., & Cohen, Y. (1989). *Item-banking of quantitative reasoning items using the 3-p logistic model* (National Institute for Testing and Evaluation Report No. 90). Jerusalem, Israel: National Institute for Testing and Evaluation.

Zach, Y., & Rahamim, I. (1999). *NITECATSYS: Man-Machine Interface (MMI)* (National Institute for Testing and Evaluation Tech. Rep. No. 94). Jerusalem, Israel: National Institute for Testing and Evaluation.

Zimowski, F. M., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *BILOG-MG – Multiple-Group IRT analysis and test maintenance for binary items*. Scientific Software International (SSI), Chicago, IL.