Economides, A. A. (2005). Computer adaptive testing quality requirements. In Griff Richards (ed.) Proceedings E-Learn 2005 World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, pp. 288-295, Vancouver, Canada, 24-28 October, AACE (Association for the Advancement of Computing in Education). ISBN: 1-880094-57-6.

Computer Adaptive Testing Quality Requirements

Prof. Anastasios A. Economides
Information Systems Department
University of Macedonia
Thessaloniki, 54006 GREECE
economid@uom.gr

Abstract: Computer Adaptive Testing (CAT) is becoming very popular. In order to be effective in the educational process, it is important to consider quality requirements. This paper presents CATE (CAT Evaluation), a framework for CAT quality requirements. It presents and analyzes three dimensions: educational, economical and technical. The educational dimension includes content, presentation, sequencing and feedback. The economical dimension includes costs, contract and cost-effectiveness. The technical dimension includes user interface, reliability, maintainability, performance, functionality, adaptation, connectivity and security. Designers, developers and evaluators of CAT systems may use this framework to make appropriate decisions.

Keywords: adaptive learning, adaptive testing, CAT, human-computer interaction, quality requirements, sequencing, usability.

I. Introduction

The quantity and quality of Computer Adaptive Testing (CAT) systems are growing rapidly, propelled by technological, standardization and psychometrics advances together with the enormous increase of examinees candidates (Straetmans and Eggen, 1998; Giouroglou and Economides, 2004). Many organizations adopt CAT (ECDL, GRE, GMAT, IMS, MS, TOEFL). Despite this delirium, little work exists on developing a common framework for evaluating the quality of CAT. Quality includes the characteristics of the system that ensure its ability to satisfy the user needs. Usability and user satisfaction are extremely important for effective CAT. Evaluation of CAT is needed to justify the investment and select the most appropriate ones. It is important to evaluate the quality of CAT in various contexts of use. For example, does the CAT accurately measure the examinee's knowledge? Does it adapt to the examinee? Is it easy to use? Is it secure? Is it cost effective?

We present CATE (CAT Evaluation), a Framework for evaluating CAT systems. For the technical characteristics of the CAT system, we are inspired by the ISO 9126 quality standard. However, we do not closely adhere to it. We extend it to best suit CAT systems. In addition, we consider the educational and economical characteristics of CAT.

The ISO/IEC 9126 standard for software evaluation defines six software quality characteristics: Functionality, Reliability, Usability, Efficiency, Maintainability, and Portability. Each of these characteristics is further decomposed in a set of sub characteristics. So, Functionality is characterized by Suitability, Accuracy, Interoperability, Compliance and Security. Reliability is characterized by Maturity, Fault Tolerance and Recoverability. Usability is characterized by Understandability, Learnability and Operability. Efficiency is characterized by Time Behavior and Resource Utilization. Maintainability is characterized by Analyzability, Changeability, Stability and Testability. Portability is characterized by Adaptability, Installability, Conformance and Replaceability.

For the technical dimension, we consider the following characteristics: 1) User Interface, 2) Reliability, 3) Maintainability, 4) Performance, 5) Functionality, 6) Connectivity, 7) Security, and 8) Adaptation. For the educational dimension, we consider the following characteristics: 1) Content, 2) Presentation, 3) Sequencing, and 4) Feedback. Finally, for the economical dimension, we consider the following characteristics: 1) Costs, 2) Contract and Licensing, and 3) Cost-Effectiveness. Each of these characteristics includes a set of parameters.

Previous studies on evaluating testing tools using specific criteria include the following. Baklavas et al. (1999) evaluate Web-based testing tools with respect to the variety of question types that support, the capabilities for multimedia use, the security, the easiness of development, maintenance and delivery of tests, the automatic grading and the statistical analysis of the results. Dunkel (1999) points out the importance of the appropriateness, reliability, validity and utility of CAT. Valenti et al. (2001) consider criteria for the interface, the question management, the test management and implementation issues. Valenti et al. (2002) suggest the use of suitability, security, interoperability, operability, understandability, learnability and reliability in order to evaluate a computer based assessment system. Sclater and Howie (2003) consider various types of users (system administrator, question author, test author, learner, marker, etc.) and propose requirements for each user type.

II. CATE Framework

In this Section, we propose CATE (CAT Evaluation), a framework for quality requirements of CAT with the following three dimensions: A) Educational, B) Economical, and C) Technical (Diagram 1).

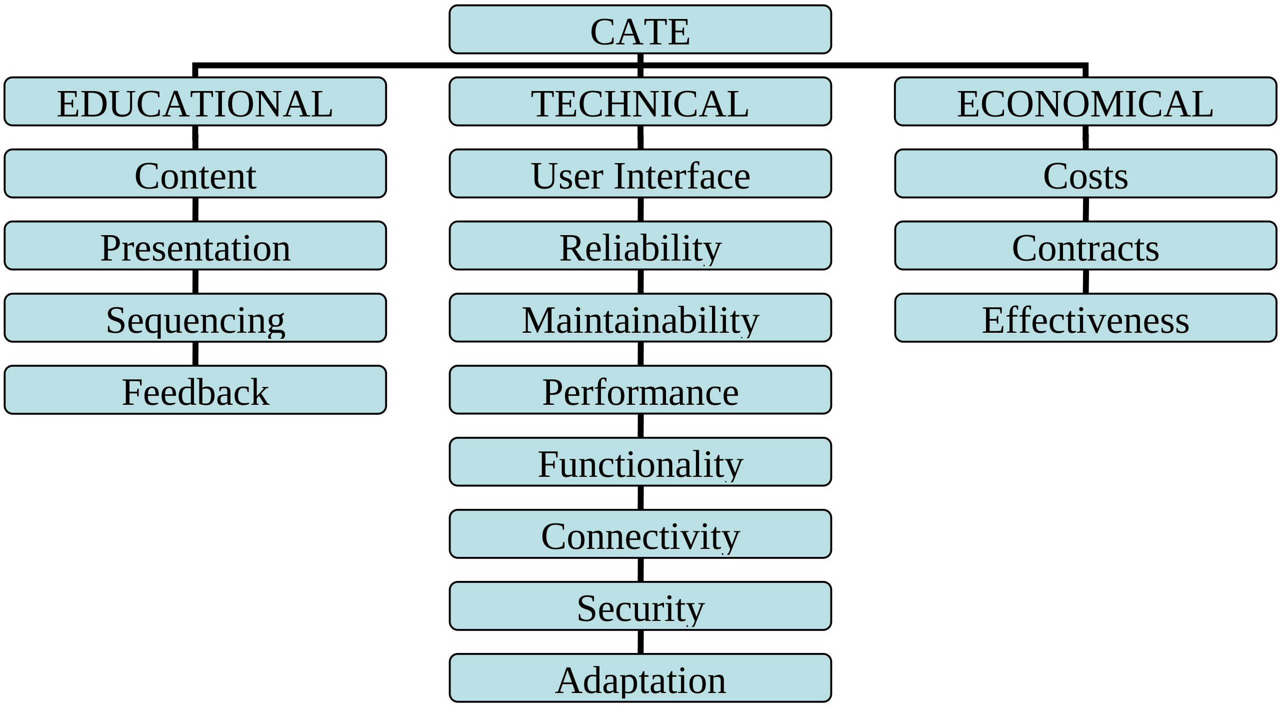

Diagram 1: CATE Quality Requirements.

A. Educational dimension

The Educational dimension includes the following: 1) Content, 2) Presentation, 3) Sequencing, and 4) Feedback (Table 1).

A1. Content

The content of CAT should be based and supported by currently acceptable didactic and pedagogical theories, such as: creative, explorative, active, constructive, problem solving, critical thinking learning. It should be personalized. The items should be of high quality, i.e. valid, trustworthy, correct and accurate without any errors. The item authors should possess credentials and reputation. The items should be useful, up-to-date, and will be valid for long time. They should be relevant, suitable and appropriate for the indented tests, ages and educational level of the examinees. They should objectively present a variety of "points of view" without discriminating with respect to age, gender, race, religious, political ideas etc. They should be acceptable and compatible to the examinee's language, social, cultural, racial, political, and religious values and ideas. They should adjust and support the values of the examinees and the value of learning.

The quantity of the items should be comprehensive and complete covering all main ideas and key points at the right quantity. It should also be sufficient and balanced to cover the intended topics, difficulty levels, skills and abilities to be tested. It should support various social interaction types (e.g. formal, informal), cognitive and conational types. Finally, it should be easy, time and cost efficient to develop, calibrate, manage, validate and update the items.

A. EDUCATIONAL	**B. ECONOMICAL**
A1. CONTENT	**B1. COSTS**
Item Personalization. Didactic Theories. Item Quality: *Validity, Authority & Trustworthiness.* *Accuracy, Authenticity & Integrity.* *Usefulness & Effectiveness.* *Uniqueness.* *Currency & Timeliness.* *Durability & Stability.* *Appropriateness, Relevance, Suitability.* *Acceptability, Multilingualism,* *Non-discriminating, Fairness, Bias-free,* *Objectivity & Variety of "Points of View".* Item Quantity *Comprehensiveness* & *Completeness.* *Appropriate and Balanced quantity.* *Variety of Interaction Types.* *Variety of Cognitive Types.* *Variety of Conational Types.* Item Management Easiness.	Planning. Buying. Operating. Maintaining. Upgrading. Terminating. Health. Environmental.
A2. PRESENTATION	**B2. CONTRACT & LICENSING**
Presentation Personalization. Media Variety. Media Quality. Clearness, Simplicity and Low Overhead. Right mix of Media. Appropriate position of objects.	Variety & Comprehensiveness. Flexibility. Duration. Visibility & Awareness. Discounts. Guarantees.
A3. SEQUENCING	**B3. COST EFFECTIVENESS**
Sequencing Personalization. Validity. Reliability. Accuracy. Consistency. Stability. Low Item Exposure. Management Easiness. Item Prioritization. Sequencing methods Variety. Scoring methods Variety. Tests and Examinees Concurrency. Guessing and Cheating Avoidance.	Examinee's Fees. Examinee's Satisfaction. Examinee's Learning. Cost-Effectiveness, Feasibility.
A4. FEEDBACK	
Feedback Personalization. Timeliness. Accuracy and Relevance. Clearness. Feedback types Variety. Quantity Appropriateness. Presentation Appropriateness.	

Table 1: CAT Educational and Economical parameters.

A2. Presentation, Media & Format

The presentation, media and format of the items should be personalized. It should be clear, simple, and of low overhead. It should be rich, be based on a variety of media (e.g. text, picture, image, graphs, diagrams, audio, video, immersion) of high quality (e.g. resolution, number of colors, sound fidelity). There should be the right mix of media objects at the appropriate positions with low distraction. The result should be enjoyable.

A3. Sequencing

The item sequencing depends on the examinee's answers. An adaptive algorithm is employed to select the next item to be presented to the examinee. This algorithm should be based on a valid and accredited pedagogical and psychometric theory. The duration and the number of items in the CAT should be enough to produce valid results. The selected items should accurately represent the content, skills and abilities that are intended to be measured. The exposure of the items should be kept low and the test-overlap minimum. The algorithm should be easy, time and cost efficient to initiate, manage and terminate. It should be fair, non-discriminating, and consistent. It should be intuitive, logical and appropriate for the examinee. There prioritization of important items. It should enhance student's motivation and enjoyment. It should support a variety of item types, sequencing methods and scoring methods. It should support a large number of concurrent tests and examinees. It should avoid guessing and cheating. It should result to valid, reliable and error-free scores. The scores should be stable, reproducibility, and consistent.

Different allocation control levels among the examinee, the teacher and the system should be possible. For example, the examinee may have the option to overtake control over the CAT ignoring any suggestions of the system. The examinee could select the next item, skip an item, go back and alter an answer, retry an item.

A4. Feedback

The feedback to the items should be personalized. It should be timely, quality, accurate, relevant, clear and easy to understand. It should be of proper quantity, media and format. It should inform the examinee about the content, the skills and abilities to be tested, the required prerequisites, the options, the available tools and resources, the CAT method and the score. It should advise the examinee on test strategies and the use of time. It should notify the examinee on deadlines. It should provide hints on the items as well explanations on the answers. It should encourage, inspire, motivate, and stimulate the examinee. Finally, it should praise and congratulate the examinee.

Support: There should exist a variety of support facilities (e.g. searching, communication, collaboration, sharing, glossary, dictionary, FAQ, bibliography, references, links, help, documentation). Also, various educational tools should be provided to the examinee and the teacher (e.g. designing, creating, and organizing the items, as well as monitoring, helping, evaluating, and recording the examinee) with no programming need. Finally, there should be a variety of communication and collaboration tools (e.g. e-mail, chat, videoconferencing, etc.).

B. Economical dimension

The Economical dimension includes the following: 1) Costs, 2) Contracts and Licensing, and 3) Cost-Effectiveness.

CAT should be economical feasible and cost-effective. The various costs should be considered together. There are costs in planning, buying, operating, maintaining, upgrading and terminating the devices, the networks, the services and the testing. There also possible health and environmental costs. There should be alternative types of contracts, for example with respect to the number of subjects, the number of tests, the number of examinees, the number of items, the test time, etc. The examinee should be aware of the various fees. The fees should be transparent at any time. The flexibility, duration, visibility, discounts (e.g. with respect to the number of tests, examinees) and guarantees are also important parameters. Finally, the cost-effectiveness is related to the achieved examinee's satisfaction, learning with respect to the fees and the costs.

C. Technical dimension

The Technical dimension includes the following: 1) User Interface, 2) Reliability, 3) Maintainability, 4) Performance, 5) Functionality, 6) Adaptation, 7) Connectivity, and 8) Security (Table 2).

C1. User Interface

The user interface should be personalized. It should be easy, time and cost efficient to understand, learn, remember and use. It should be simple and convenient to use (e.g. minimum number of clicks to find and display information, minimum number of scrolls to display information). It should facilitate both test authoring and taking. It should support the examinee's focus and attention, avoiding her distraction, boring and tiredness due to cognitive load. It should consider examinees with disabilities and do not discriminate. It should treat all fairly.

Its features and operation should be appropriate, convenient, meaningful, self-evident, and rational. It should be uniform and consistent. Under the same conditions similar results should be produced (e.g. messages, colors, menus). Its operation should be correct, accurate, precise and effective.
Its layout, organization and structure (e.g. frames, menus, buttons) should be simple, intuitive, rational and effective. Its design should be aesthetic, attractive, pleasant and fun to use it. It should support many languages and media types (e.g. text, audio, video, immersion) of high fidelity at the right mix and position on the user interface.
Its navigation should be easy, simple, intuitive and rational. There should be alternative ways of navigation with proper number of levels. It should offer many navigation facilities (e.g. sitemap, next, previous, home, exit, undo, redo, shortcuts, history, save, print). Orientation & Help: It should provide quality orientation and help (e.g. documentation dictionaries, FAQ, search engine) in a consistent way. It should provide feedback (e.g. error messages) and messaging (e.g. notifying, alerting for deadlines, events) in various format, presentation and media.

C. TECHNICAL	
C1. USER INTERFACE	**C4. PERFORMANCE**
Personalization. Easiness: *Understandability.* *Learnability & Familiarization.* *Rememberability.* *Usability, Operability, Productivity & Ergonomics.* *Simplicity & Convenience.* *Attention, Focus, No Distraction & Cognitive overhead.* *Multilingualism,* *Accessibility, Non Discriminating, Bias-free & Fairness.* Quality: *Consistency & Uniformity.* *Appropriateness, Meaningfulness, Self-evidence, Rationality*	Processing Speed. Communication Bandwidth. Memory Capacity. Energy Consumption. Responsiveness, Delay & Timeliness. Input. Output. Measurability. Controllability. Evaluability. Effectiveness, Efficiency & Resource Utilization.
& Causality.	**C5. FUNCTIONALITY**
Correctness, Accuracy & Precision. *Effectiveness.* Layout, Organization & Structure: *Simplicity.* *Intuitiveness &Rationality.* *Number of Levels & Choices per Level.* *Aesthetics, Attractiveness, Pleasance, Enjoyability & Fun.* Media: *Variety & Comprehensiveness.* *Quality & Fidelity.* *Appropriateness, Right Quantity, Mix & Positioning.* Navigability: *Easiness, Simplicity, Intuitiveness & Rationality.* *Flexibility & Variety of Alternatives.*	Comprehensiveness, Completeness, Variety of Features & Applications. Quality. Simplicity, Self-explanatory, Intuitiveness & Rationality. Usefulness, Suitability & Productivity. Accuracy. Timeliness. Synchronization, Coordination, Concurrency & No Interference. Autonomy & Self-contained. State-of-the-art, Currency & Innovativeness. Maturity & Stability.
Right Number of Levels.	**C6. ADAPTATION**
Variety of navigation facilities. Interactivity & Feedback: *Multimedia Communication.* *Variety & Comprehensiveness.* *Quality & Fidelity.* *Responsiveness, Timeliness & Synchronization* *Appropriateness, Right Quantity, Suitability & Usefulness.* *Consistency.* Orientation & Help: *Variety & Comprehensiveness.* *Quality.* *Appropriateness & Right Quantity.* *Consistency (e.g. use of terms).*	Comprehensiveness & Variety: Educational parameters. Technological parameters. Economical parameters. Transparency. Correctness, Accuracy & Precision. Usefulness, Appropriateness & Effectiveness. Timeliness & Speed of adaptation. Consistency of adaptation. Flexibility & Adjustability. Prioritization of parameters.

C2. RELIABILITY	**C7. CONNECTIVITY**
Error Free. Error Prevention. Monitorability, Analyzability, Testability, Error Recognition & Error Diagnosis. Fault Tolerance, Robustness, Troubleshooting, Error Recovery, Recoverability, Resumability & Error Transparency. Availability. Stability, Maturity & No Volatility. Correctness, Accurateness & Precision. Consistency. Back-up & Mirroring. Documentation, FAQ & Help. Certification, Guarantees, Brand Name & Reputation.	Openness, Compliance and Conformance to Standards, Compatibility, Interoperability & Universality. Portability, Exportability, Diffusiability, & Reusability. Transparency, Seamless Integration, & Harmonious Interconnectivity. Scalability, Extensibility & Expandability. Comprehensiveness & Variety: *Platforms* *Interfaces* *Multimedia Format* *Item Types* *Databases* *Testing Methods*.
C3. MAINTAINABILITY	**C8. SECURITY**
Easiness & Flexibility of Maintenance. Installability & Easy of Installation Reconfigurability, Self-Tuning, Modifiability & Changeability. Reparability & Replaceability. Integrity & Survivability. Improvability, Upgradeability & Revision easiness. Supportability.	Comprehensiveness of Security Technologies: *Firewalls.* *Access Control, Authorization & Authentication.* *Certification.* *Encryption & Cryptography.* *Tunneling.* *Anti-Virus, Anti-Spa, Anti-Spy*. Levels of security. Cheating Prevention. Confidentiality, Privacy & Anonymity. Trust. Control of Personal Data by examinee. Certifications & Guarantees.

Table 2: CAT Technical parameters.

It should support a variety of rich and of high quality interactivity and multimedia communication (e.g. one-to-one, one-to-many, many-to-many, synchronous, asynchronous). The interactivity and the multimedia communication should be at the right quantity at the right moment without producing cognitive overload. The responses to any examinee's action should be immediate and effective.

C2. Reliability

Reliability is related on the capability of the CAT to maintain its level of performance under stated conditions for a stated period of time. The CAT system should be error-free. It should prevent errors that may occur, for example measurement errors. It should be easy and fast to be monitored and tested. If an error or fault happens, it should recognize its existence and its source. It should make correct diagnosis of the error. The error should be easily repaired by the system or by external intervention with minimum effort and resources at the minimum time. No data or other useful resources should be lost in case of error. The repair should be transparent to the examinees. No data discrepancies should occur due to hardware faults (e.g. power off, communication disconnection). The duration and the cost of the interruption should be minimal. The CAT should handle any unexpected case and should resist to malicious attacks. It should not be stacked in a deadlock situation. Its operation should be stable and consistent with minimal transient phenomena. It should always be available.

Its operation should be correct and accurate. It should do what is supposed to do, for example alerting examinees about deadlines. Its operation should be consistent and similar states should be treated similarly. For example, examinees at the same performance level should be taken assessments at the same difficulty level. It should keep on back of all data, items, scores, statistics, etc. The perceived reliability of the system increases with the reputation and the brand name of the manufacturer, as well as with awards, certifications and guarantees that are given to it.

C3. Maintainability

Maintainability is related on the effort needed to maintain the CAT and make specific modifications. Initially, the installation of the CAT should be easy and fast. The CAT should need minimal effort and time to maintain its efficient operation. In case of changes in the CAT's scope and operation, its reconfiguration should be easy, unproblematic and fast. In case of faults, the repair or replace of the faulty parts should be fast and easy. It should be easy and fast to be revised and upgraded. Its integrity, resistance and survival from attacks should be guaranteed. Its efficient operation should be supported by the manufacturer. The guarantees should be for long time and take care of any possible case.

C4. Performance

Performance is related on the achieved performance and efficiency of the CAT. The processing speed should be high enough to efficiently operate it. The communication bandwidth (both for uploading and downloading) should be high enough to support any possible communication. The memory capacity should be large enough to store all possible data, items, etc. The energy consumption should be small enough. The response of the system to a change (e.g. examinee's response) should be fast and appropriate. The delay of processing, storing, communicating data should be smaller than the threshold for efficient CAT. The quality and the fidelity of the input (e.g. camera, handwritten recognizer, speech recognizer) and output (e.g. screen, speakers) should be appropriate. For example, the quality of the displayed, stored and transmitted images should be the best possible given the constraints (bandwidth, delay etc.). So, the camera and screen resolution, the screen size, the ergonomic keyboard are important factors. The performance should be easily measured, evaluated and controlled. Finally, the effectiveness and efficiency of the system should be high.

C5. Functionality

Functionality is related on the available functions, features, tools, and applications in the CAT. Examples of tools include: editor, drawing, audio recorder, photo camera, video recorder, fingerprint reader, handwriting recognition, speech recognition, face recognition, multimedia processing, etc. Examples of features and applications include: multimedia mail, alerting and reminding, chat, telephony, videoconference, etc. These features and applications should be of high quality, simple, self-explanatory, intuitive and rational to use them.

Each feature or application should function autonomously and be self-contained. There should be no need for extra plug-ins. Multiple features and applications should function concurrently synchronized with no interference among them. The technology used to implement the system should be not only current and innovative, but also mature and stable.

C6. Adaptation

CAT should adapt its educational parameters (e.g. content, presentation, sequencing, and feedback), its technological parameters (e.g. user interface, security), and its economical parameters to the examinee and the teacher. CAT should be personalized. For example, it should adapt the next item to be presented to the examinee according to her answer. It should adapt the content to the screen size. It should adapt the resolution of an image item to the available transmission bandwidth.

The adaptations should be transparent to the examinee. They should be correct, accurate, precise, and error free. They should be useful, appropriate and effective. They should also be timely. They should be consistent and uniform, similar results should appear for similar reasons. They should be flexible and adjustable, i.e. if an exact match cannot be found an approximation should be given. Also, there should be prioritization among the parameters importance in case of constraints or conflicts.

C7. Connectivity

Connectivity is related on the ability of the CAT to be connected to other software and hardware systems. The CAT system should provide as much connectivity (inside and outside of the system) as possible. CAT, items, tools, resources, examinees and teachers should be smoothly interconnected. It should follow open architectures, comply with international standards and be compatible to as many software and hardware devices as possible. It should easily import and export data, items, scores, statistics, etc. All parts should be seamlessly integrated to construct the whole CAT. The integration of the parts should be transparent to the examinee. All interconnections should be done in harmony with minimum examinee's effort. It should be easy and fast to add or remove multiple items and tests. Also, as many as possible items should be reused by multiple tests. These reusable items would be also exported and be used by other CATs. Also, it should be easy and fast to connect or disconnect as many concurrent examinees as possible. It should support multiple platforms, databases, item types, multimedia format, testing algorithms, etc. Finally, it should be autonomous not required additional plug-ins.

C8. Security

CAT should support current, updated security technologies (e.g. firewalls, access control, authorization, authentication, certification, encryption, cryptography, tunneling, anti-virus, anti-spam, anti-spy) to protect the items, the adaptive testing algorithm, the examinees' data, the scores, etc.

It should protect both the storage and the communications. It should support multiple levels of security for different examinees and resources. It should prevent cheating, plagiarism, unauthorized notes taking, reproduction and coping, communication and collaboration, access to data, tools and resources, overtime access to test, etc.

It should support the examinee's confidentiality, anonymity, privacy and trust. The examinee should have control of what personal information should be available to others. There should be no secret activities occurring. All data, activities, decisions and applications should be visible and available to the examinee whenever she requests them. For example, there should be no secret monitoring and recording of the examinee's transactions. High prestige security organizations should certify and guarantee its security.

III. Conclusions

Every effort should be made to support the examinee. The examinee should take the appropriate exam accurately, being satisfied and using the resources efficiently. We provide insights on the user requirements of CAT. We propose CATE (CAT Evaluation), a framework for quality issues of CAT systems. Designers, developers and evaluators of CAT should consider educational, economical and technical characteristics. We analyze these dimensions and suggest guidelines for design, development and evaluation of CAT. We believe that the CATE framework will not only sensitize readers to relevant parameters, but also provide researchers with a map that can help motivate studies on this topic. In a forthcoming paper, we evaluate several CAT systems using CATE.

Acknowledgments

The work presented in this paper is partially funded by the General Secretariat for Research and Technology, Hellenic Republic, through the E-Learning, EL-51, FlexLearn project.

References

Baklavas, G., Economides, A.A., and Roumeliotis, M. (1999). Evaluation and comparison of Web-based testing tools. In *Proceedings WebNet-99, World Conference on WWW and Internet*, pp. 81-86, AACE 1999.

Dunkel, P. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning & Technology*, Vol. 2, No. 2, January 1999, pp. 77-93.

Giouroglou, H., and Economides, A. (2004). State-of-the-art and adaptive open-closed items in adaptive foreign language assessment. In *Proceedings 4th Hellenic Conference with International Participation: Informational and Communication Technologies in Education*, Athens, pp. 747-756.

ISO/IEC 9126: Information technology - Software Product Evaluation - Quality characteristics and guidelines for their use - 1991. http://www.iso.org

Sclater, N., and Howie, K. (2003). User requirements of the ultimate online assessment engine. *Computers & Education*, 40, pp. 285-306.

Straetmans, G.J.M., and Eggen T.J.H.M. (1998). Computerized adaptive testing: what it is and how it works. *Educational Technology*, January-February 1998, pp. 82-89.

Valenti, S., Cucchiarelli, Al, and Panti, M. (2001). A framework for the evaluation of test management systems. *Current Issues in Education*, 4, 6.

Valenti, S., Cucchiarelli, Al, and Panti, M. (2002). Computer based assessment systems evaluation via the ISO9126 quality model. *Journal of Information Technology Education*, Vol. 1, No. 3.