Giouroglou, H. & Economides, A.A.: **The development of the adaptive item language assessment (AILA) for mixed-ability students.** *Proceedings E-Learn 2005 World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, AACE, 2005.

The Development of the Adaptive Item Language Assessment (AILA) for Mixed-Ability Students

Giouroglou, H.
University of Macedonia
Greece
hara@uom.gr

Dr. Economides, A.
University of Macedonia
Greece
economid@uom.gr

Abstract: Cross-cultural education has created diverse, mixed-ability students. The one-to-many, teacher-centered tutoring model that was used in traditional education up to date and in traditional education and the previous generation of Computer Aided Instruction (CAI) is no longer applicable to distance education due to students' heterogeneity. In order to foster learners' success, we need to adapt Foreign Language Assessment (FLA) environments to accommodate learners' diversity accordingly. Computer Adaptive Testing (CAT) is a promising field under research that can create FLA adapted to individual student needs, abilities, backgrounds, strengths and weaknesses, giving emphasis to cognitive language skills, such as comprehension, production and use. This paper will analyze the current trends in CAT for FLA and it will focus on the inability of modern systems to accommodate students' diversity. Finally, it will describe the development of a computer adaptive placement test for mixed-ability students that can measure both the breath and depth of foreign language awareness in little time.

Introduction

The promotion of multilingualism, communication, mobility and cross-cultural awareness among the EU and globally has created the demand for easily-administered, self-paced, time-effective, multilingual, and internationally accredited foreign language assessments (FLA). Students of all age groups participate in training and lifelong-learning programmes matching both their occupational and personal needs, due to the demand for the constant retraining of the global workforce [Twigg, 1994]. Modern education needs to provide the new student society with the tools to construct their own knowledge with their own pace, ability, individual learner characteristics and aptitude [Schunk, 1996]. The Common European Framework of Reference (CEF) has paved the way by setting internationally certified standards for formal as well as self-assessment in all European languages. CEF has created six clearly defined proficiency levels (A1,2/B1,2/C1,2) for both formal and self-assessment purposes. The levels describe what receptive and productive skills the examinee needs to possess in order to attain the desired level of competence [Council of Europe, 2001]. There is a perceived need for a new generation of FL tests, which should be adaptive and adaptable in nature, catering for diverse, mixed-ability students. Students' diverse needs and abilities pose the necessity for the development of flexible assessments that will suit the diverse student's cognitive skills. Modern FLA implementations fail to cater for mixed-ability students, as they are linear and targeted to the average student. Computer Adaptive Language Testing (CALT) technology can provide student-centered assessment, replacing traditional testing wherever possible. This paper will describe the new trends in CALT and will analyze the Adaptive Item Language Assessment (AILA), a placement CEF test assessing competence in English language.

State-of-the-Art in CAT

The major advantage of CAT systems is that they are student-centered as, in contrast to their paper-and-pencil (P&P) counterpart, they can be tailored to the ability and level of each examinee by updating the estimate of the examinee's ability, and adapting the subsequent items to the individual ability of each examinee. Item adaptation results in reduced standard errors and improved accuracy of scores for both high and low ability test takers. Tailored item selection also leads in avoidance of examinees' boredom from answering too easy questions and of frustration from answering too difficult questions. Thus, CAT is said to have increased efficiency, greater precision with less items, and time-effectiveness, since only a few tailored items are needed to achieve accuracy. CAT systems offer also greater test security and longer duration [Wainer et al, 2000] than traditional P&P tests, as they are comprised by large item pools with controlled item exposure, rendering examinees incapable of knowing the items in advance. CAT shares all advantages of CBT, such as immediate feedback and self-pacing.

CALT uses adaptive technologies to assess foreign language (FL) competence. Most international FL testing organizations have started delivering their tests in self-paced CAT mode, making their tests available to even more people. CALT successfully assesses multiple-choice (MC) items in vocabulary, grammar, reading and listening, using the Item Response Theory (IRT) [Laufer, et al, 2001, Dunkel, 1997]. Yet, CALT systems do not have adaptive components in oral proficiency and writing tasks. Open writing tasks can be marked by electronic marking, such as the Intelligent Essay Assessor (IEA) and e-rater of GMAT – with the aid of human markers [Streeter et al, 2002, Powers, 2001]. As such, CAT is not applicable to all subjects and skills, as it is based on the IRT model, which is not applicable to all item types. To achieve accuracy, IRT requires careful item calibration, excluding items that cannot be easily calibrated, such as open-ended questions [Lord, 1980]. Another crucial drawback is that the examinees are not permitted to go back and change answers, as the program selects next item on the basis of the previously answered item(s). This renders reviewing implausible, and in many cases examinees that sat both P&P and CAT failed or achieved low marks in computerized testing. To sum up, CAT systems cannot specialize on every plausible item. Although IRT increases the validity and reliability of the test, it lacks the flexibility to cover a wide range of activities and abilities, including open answers, and productive language use.

The Problem

In terms of foreign language assessment (FLA), research in neuropsychology and especially psycholinguistics has revealed that individuals process language differently, according to their overall intelligence, brain dominance, sex, inherent traits and cognitive skills [Akmajian, et al. 1998]. There is an apparent relationship between language, thought and cognition as Chomsky and Piaget have advocated from different points of view [Chomsky, 1997]. Accordingly, educationalists acknowledge the fact that there are mixed-intelligence students, meaning that learners can attain new knowledge using different learning strategies and paths, suited to their individual intelligence [Gardner, 1993]. Finally, research in first language acquisition has revealed serious findings in human language development through the study of certain phenomena, such as hesitations, speech errors and language disorders that can also be applied in second/foreign language acquisition.

However, CALT nowadays is based on solid programming that is collective rather than individualized and fails to include crucial cognitive parameters of student language competence and performance. Such systems cannot replace the human examiner without nasty consequences for its group of examinees. The new generation of assessment systems for cross-cultural examinees should not assess students horizontically as an equable lot but vertically as mixed-ability individuals with mixed-scoring options. Moreover, the new generation of assessment should create different experiences that will motivate and exploit the different skills of the test-takers [Ali, 2001].

The majority of CALT systems use MC items to discriminate among proficient, good and weak learners. This is mainly due to the fact that MC items are easily programmed and calibrated in IRT. The program can easily identify correct and wrong answers and move on to easier or more difficult items. This technique is also reliable and valid as long as items are adequately pre-tested and correctly calibrated. However, MC items cannot allow active expression and language production. Examinees are passive viewers of the proposed answers and they only try to segregate the correct answer out of the distracters. Proficient learners answering MC items are not given the opportunity to discriminate themselves from other learners by openly typing the correct answer in case they know it. They have to choose among the four intended choices and receive the same mark as other learners who will purposefully or accidentally choose the correct item. This limitation does not allow the proficient learner discern from others by testifying active language production. Another problem is caused by the prohibition of item reviewing. In psycholinguistics there is a clear discrimination between errors, made due to ignorance, and mistakes, made due to negligence. Examinees are prone to mistakes not only out of ignorance but also out of misunderstanding, anxiety, confusion, distraction or other physical reasons. Since reviewing is impossible, adaptive systems may form false impressions and give low scores. To this end, CALT should become more "intelligent" and simulate the human examiner in order to be more accurate and precise in their scores.

AILA : A CEF Placement CAT, System Architecture

The Adaptive Item Language Assessment (AILA), developed by CONTA Lab at the University of Macedonia, is an adaptive placement test, based on CEF standards, that is both adaptive and adaptable in that examinees are given the choice to select how to answer each item presented. Examinees can choose between two options, the first in MC and the second in open-answer (OA) mode. The system adopts course content tailored to the student's needs, taking into account different difficulty as well as different knowledge levels.

It is important to create a system that is affordable, and easily maintained. To achieve this, it is required to create re-usable objects of previous existing educational content. We used a CPU Pentium III 800 MHz, with 2 GB RAM, and the Apache web server. The software can run on Windows NT 4.0, Windows 2000 Server or Advanced Server. The system software includes the required MySQL database software. For reasons of re-usability XML has been used to separate content from the way it is processed (i.e. presented) and which avoids to re-write the same content that needs to be displayed in different formats. The software used is Windows 2000, My SQL (free), PHP, VB script, Javascript, HTML, XML. The system has a modular, component-based architecture that makes it easy to create the adaptive testing system and to re-use data from different learning levels. It is an independent platform and avoids vendor lock-in.

System Description

AILA [fig. 1] measures the ability of non-native speakers of English to use and understand English as a Foreign Language for achievement and placement purposes. The test-takers who sit AILA can quickly assess their competence in English in the scale issued by the Common European Framework of Reference (CEF) for foreign language assessment. The Bulletin Board and the F.A.Q. components constitute this application a complete web-based, self-access FLA. The test measures competence in four out of the six CEF levels, A2, B1, B2 and C1 each of them consisting of three item types:

- Grammar and Structure (Use of English) items. They measure the ability to recognize and/or produce grammatically English Language structures that are appropriate for each CEF level.
- Vocabulary (Use of English) items. They measure the ability to recognize and/or produce high or low frequency English words that are appropriate for each CEF level.
- Reading items. They measures the ability to understand and extract information from short passages that are appropriate for each CEF level.

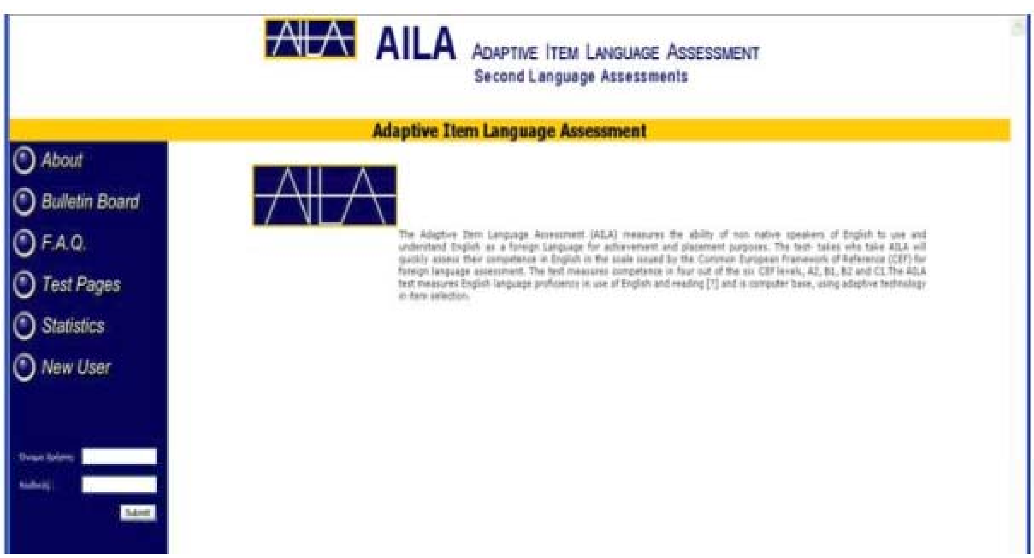


Fig. 1: AILA User interface

Content Adaptation

AILA measures English language proficiency in Use of English and reading and is computer-based, using adaptive technology in item selection. The system increases student motivation, by providing tailored content adapted to his/her needs and level of competence. This also results in a reduction of time spent with maximum benefit. To create an open and flexible testing environment, the student defines his/her educational goals by choosing his knowledge level according to which the test will be administered in a predefined way.

As the test is administered, examinees are given the freedom to choose between two options [Chart 1]. They can either type an open answer (OA) or choose the correct answer in MC mode. If the examinee knows the answer and is able to produce it, then he/she can type the answer in the OA mode and has the opportunity to demonstrate his/her advanced knowledge. A correct OA response receives a bonus in the total score (grade+ 0.25) and updates the User Profile of the examinee. Then, the item selection algorithm proceeds to the next item of increased difficulty. A wrong OA response does not affect the final score and it immediately directs the examinee to the MC mode of the same item. When the MC mode appears, the examinee cannot go back to the OA mode.

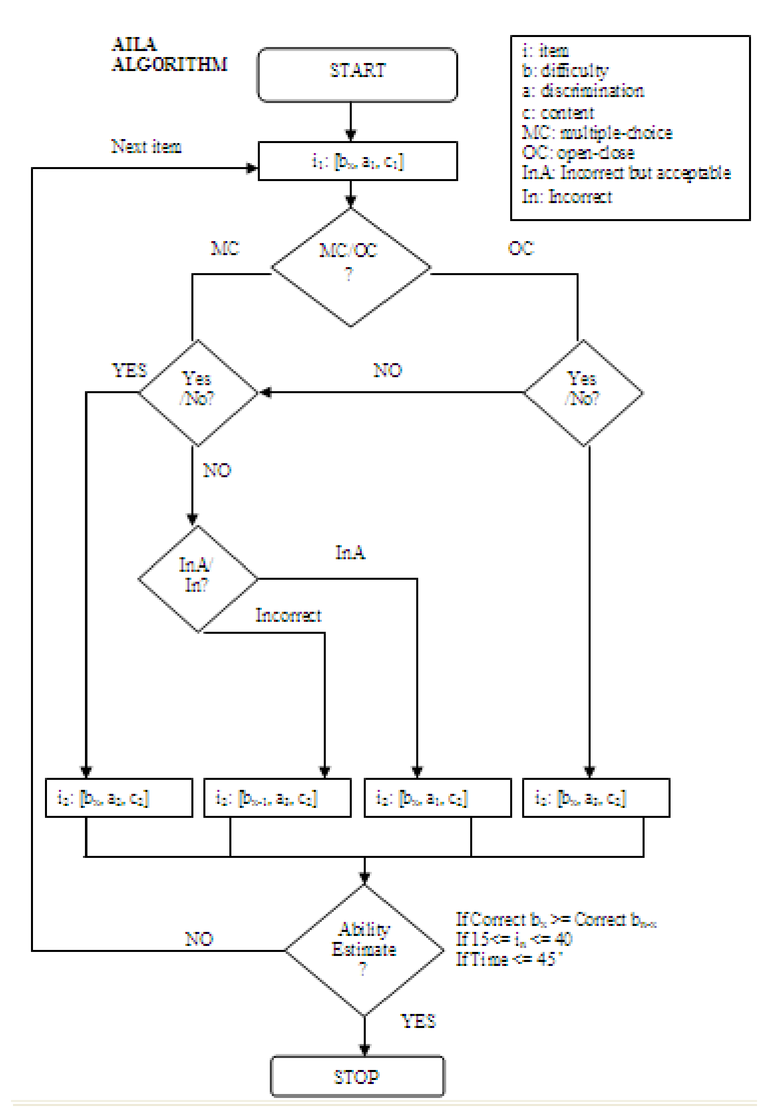


Chart 1: AILA flowchart

The immediate selection of the MC mode does not have a negative effect on the score, as correct MC choices receive the highest mark (1), and the adaptive algorithm immediately proceeds to the next, increased difficulty item. Wrong choices receive no mark and the next item is easier. This method does not affect the final score of the test or punish a wrong OA answer. Instead, it gives the opportunity to the examinees to demonstrate productive FL use and active FL extraction from their long-term memory.

The duality of the system enables the test-taker to have many answering options and various evaluations. In order to cope with students' divergent cognitive strengths and weaknesses, AILA tries to discern between errors and mistakes, using a simple method. The MC destructor that bears a close resemblance to the correct option is be regarded "acceptable" and the item selection algorithm proceeds to an item of equal difficulty. The reason for doing this is the fact that in most MC questions at least one destructor is so close in meaning or in grammatical resemblance to the correct answer that may sometimes puzzle even examiners. Bearing in mind the fact that language is a flexible, ever-changing, living entity used to communicate meaning and retrieve information, we should create CATs that will accept answers that have a slight deviation from the standard form. It is also a fact that while native speakers of every language tend to do mistakes in oral and written language production, they are still fluent and proficient speakers of their mother tongue.

The Learner Model

The LM reflects specific characteristics of the learner and thus it is used as the main source of the adaptive behavior of AILA. The information held is divided into domain dependent information and domain independent information. As far as the domain dependent information is concerned, the LM keeps information about: (i) the learner's knowledge level (qualitative – which levels of competence – and quantitative – how many items are correct – estimation) with respect to the average level of the items answered correctly, (ii) the learner's errors, and (iii) the learner's behaviour during his/her interaction with the tool in terms of the frequency of errors made, time of response, etc. As far as the domain independent information is concerned, the LM keeps general information about the learner such as username, age, sex, learner's right or left-handedness, last time/date the learner logged on/off. The LM is dynamically updated during the learner's interaction with AILA in order to keep track of the learner's "current state".

Item Bank and Stopping Rule

The item bank [fig.2] consists of 600 items divided in the four CEF levels of competence (A2, B1, B2, C1), signifying item difficulty ($b_{1\text{-}4}$). In each broad level of competence, items are sub-divided in three discrimination levels ($a_{1\text{-}3}$): The first discrimination level (a_1) contains items that are expected to be answered correctly by all examinees having the given competence, the second (a_2) contains items that can be answered correctly by the average examinee, while the items in the third level (a_3) can only be answered by the most competent students in this level. Finally, each discrimination level is seperated in 5 content areas ($c_{1\text{-}5}$), in order to ensure that examinees will answer a wide variety of language items.

The test starts with a given difficulty specified by the test-taker (b_x), low discrimination (a1), first content area (c1), and random item selection. If the test-taker answers in MC mode correctly, then the next item is of the same difficulty (b_x), medium discrimination (a2), second content area (c2), and random item selection, otherwise the next item is one difficulty level lower. If the examinee answers in OC mode correctly, then the next item is of higher difficulty (b_{x+1}), high discrimination (a3), fourth content area (c4), and random item selection. With this stratified way, we ensure that examinees will gradually attain their level of competence, by answering different item types. The stopping criterion could be time, number of items administered, change in ability estimate, content coverage, a precision indicator such as the standard error, or a combination of factors. In a variable-length adaptive test, the number of items administered to each examinee differs depending on the number of correct/incorrect responses given by him or her to the items presented. A variable-length stopping rule terminates a test once a pre-specified level of measurement precision has been reached, based on the standard error associated with a given ability. The advantage of implementing variable length stopping rules is that all examinees' ability estimates have the same measure of precision [Thissen and Mislevy, 2000].

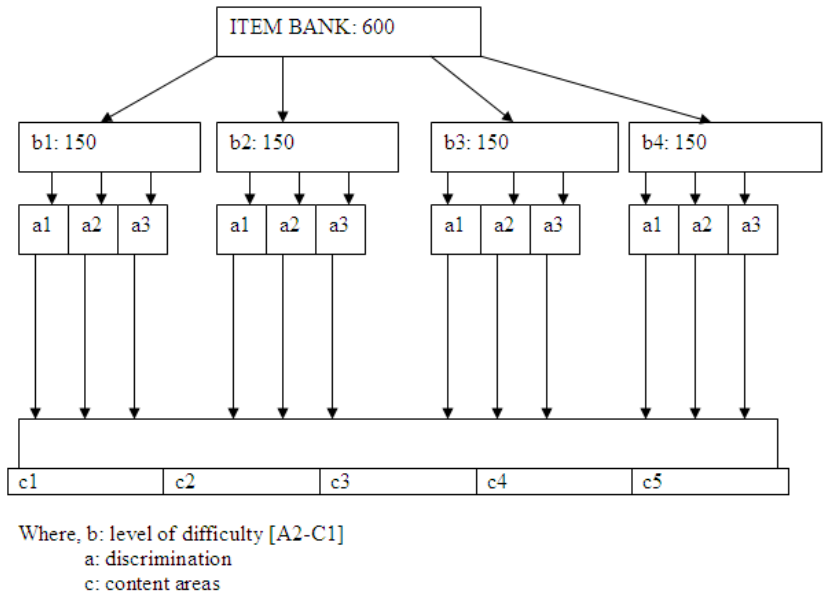


Fig.2: Item bank architecture

However, a non fixed-length stopping rule has the potential to produce adaptive tests that are much shorter than P&P tests and this may have a negative effect on examinee reactions and scores. Therefore, AILA algorithm has a compulsory minimum number of 15 required items [Chart 2]. Thus, the minimum test length is 15 items and the maximum is 40 items. The test stops when the examinee answers at least 15 items, having shown competence at one level of difficulty. There are no time limits per item; however, the maximum test time is 45 minutes.

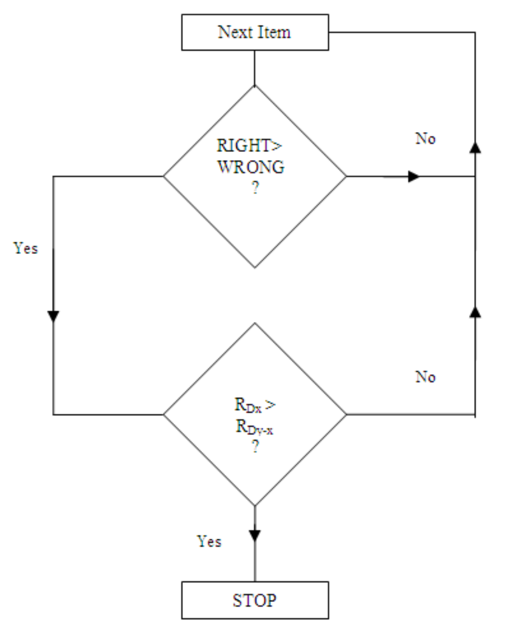
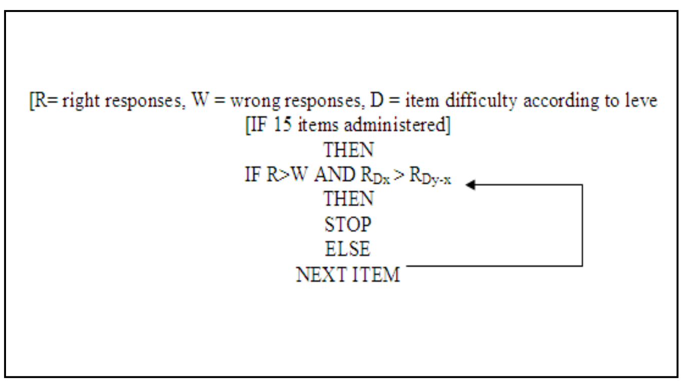


Chart 2: AILA stopping rule



Performance Information /Statistics

Statistical analysis of test scores gives useful information about the performance of individual test items, as well as the performance of mixed-ability students regarding each item. The items used in AILA are pre-tested in P&P form in order to be calibrated in terms of their level of difficulty and their discrimination. This kind of analysis provides information about item facility/difficulty, that is the proportion of correct responses to the item, item discrimination, and distractor tallies. The distractors of a MC item do not function equally, as statistical analysis proves that some of them might attract a large number of candidates. This information facilitates the development of a reliable item pool for AILA. Statistical item and student analysis also takes place during the CALT and the results are viewable after the administration of each test. The statistical information in AILA is both collective and individual.

Collective analysis includes the number of candidates and items, the score variance – the spread of scores around the mean score –, and the minimum/maximun/median score. It also provides information about mix-ability students, by measuring the performance of left/right-handed, male/female [fig. 3], and productive/receptive candidates. Individual analysis shows the performance of each candidate as well as the history of his/her earlier administrations. Individual candidate performance shows all the items administered in the specific session and how they were answered. Thus, the analysis shows the level, as well as the percentage of the candidates productive and receptive competence.

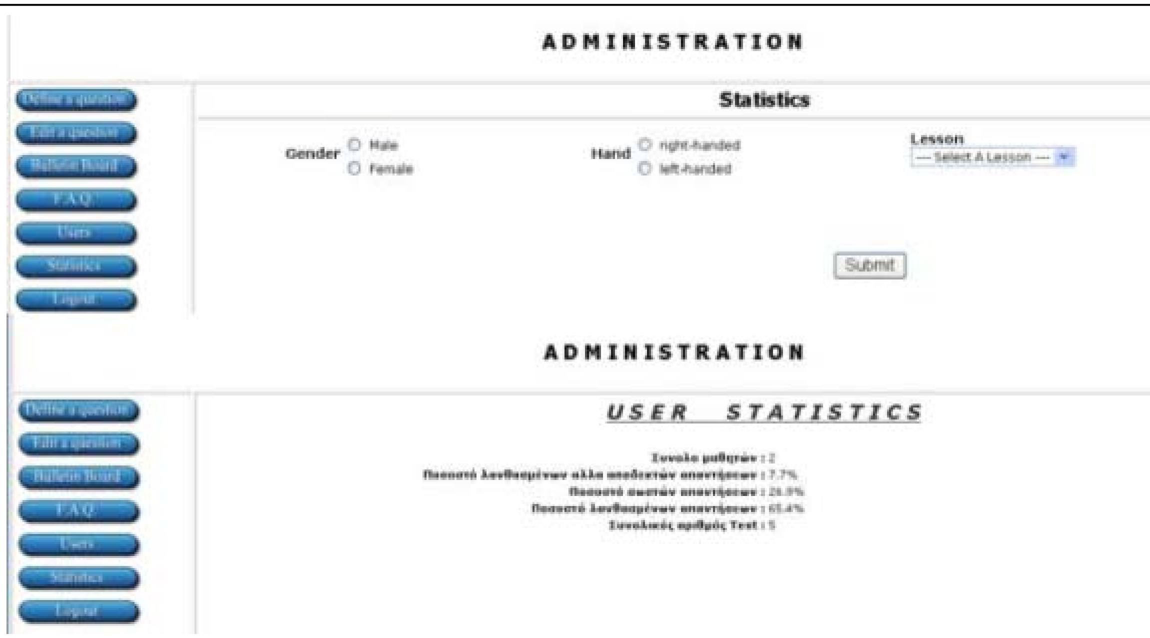


Fig. 3: AILA statistics

Outcomes and Conclusion

Modern learning societies consist of cross-cultural, mixed-ability and mixed-intelligence students. FLA tools need to adapt to this challenge, using adaptive technologies to provide personalized, self-paced and time-effective assessments. CAT can introduce a new, student-based era in FLA that will be personalized, flexible, and sensitive to human cognition, language processing and error correction. To this end, we developed AILA, an adaptive placement test that measures competence in EFL in terms of CEF levels, giving students the choice to show productive and receptive language use. The system also tries to discern errors from mistakes by evaluating students' answers. Thus, proficient learners will be able to excel, showing active language production. All in all, CALT needs to adapt to the new social conditions, adopting a new test theory [Mislevy, 1996], gathering information from various disciplines and assimilating this information in the new assessment medium.

References

Akmajian, A. et al. (1998) *Linguistics. An Introduction to Language and Communication.* Third Edition. The MIT Press.

Ali, A. (2001) "Technology Integration and Classroom Dynamics" In *Proceedings of AACE Webnet 2001 World Conference on the WWW and the Internet,* October 23-27, Orlando, Florida, USA, pp. 7-8.

Burstein, J. and Chodorow, M. "Automated essay scoring for nonnative English speakers." *In Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing*. June 1999, College Park, MD.

Burstein, J. et al. "Using lexical semantic techniques to classify free-responses." In *Proceedings of the ACL SIGLEX Workshop on Breadth and Depth of Semantic Lexicons.* June 1996, Santa Cruz, CA: University of California, Santa Cruz.

Chomsky, N. (1997) Powers and Prospects. Reflections on Human Language and the Social Order. Second Edition. Pluto Press.

Council of Europe, (2001) *Common European Framework of Reference for Languages***.** Cambridge University Press.

Dunkel, P. "Computer-Adaptive Testing of Listening Comprehension: A Blueprint for CAT Development" *The Language Teacher Online*, October 1997.

Gardner, H. (1993) *Frames of Mind: The Theory of Multiple Intelligences.* Basic Books.

Laufer, B. and Y. Yano. 2001. Understanding unfamiliar words in a text: do L2 learners understand how much they don't understand. *Reading in a Foreign Language* 13: 549-566.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mislevy, R.J. (1996), "Test Theory Reconceived." *Journal of Educational Measurement*, 33 (4), 379-416.

Powers, D. E., et al. (2001). *Stumping e-rater:® Challenging the validity of automated essay scoring* (GRE No. 98-08bP, ETS RR-01-03). Princeton, NJ: Educational Testing Service.

Schunk, D. (1996), *Learining Theories: An Educational Perspective*. Pentice Hall.

Streeter, L. et al. (2002), "The Credible Grading Machine: Automated Essay Scoring in the DoD" Paper presented at the Interservice/Industry, Simulation and Education Conference (I/ITSEC). December 2-5, 2002. Orlando, FL.

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. H. Wainer (Ed.), *Computer Adaptive Testing: A Primer* (2nd ed., pp. 101-133). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Twigg, C. (1994), "The Need for a National Learning Infrastructure", *Educom Review*, 29, Nos. 4,5 and 6.

Wainer, H. Et al. (2000). *Computer Adaptive Testing: A Primer* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.