State-of-the-Art and Adaptive Open-Closed Items in Adaptive Foreign Language Assessment

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

A. Economides
Associate Professor, University of Macedonia
Thessaloniki, Greece
economid@uom.gr

SUMMARY

In the era of multilingualism, foreign language testing needs to be immediate and accurate. This paper will make a full description of the state of the art in Computerized Adaptive Language Testing, with reference to current implementations in this field. Moreover, the paper will focus in one of the main problems in CALT, regarding the valid and reliable discrimination between the proficient foreign language (FL) examinee and the good examinee. Finally, the paper will describe a method applicable in close-ended items that will give the opportunity to proficient FL examinees achieve higher marks without influencing the scores of the other examinees.

KEY WORDS: *Computer Adaptive Testing, English as a Foreign Language, multiple-choice*

INTRODUCTION

Computer Adaptive Language Testing (CALT) is a recent development in personalised learning. Research in this area aims to create systems that will measure language proficiency as accurately as traditional means of foreign language testing. Tomorrow's 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). Also, the rapidly evolving Information Society demands constant retraining of the global workforce (Twigg, 1994), resulting to the need of easily administered, and valid assessment tools. Computer Adapted Testing (CAT) promotes personalised learning, as it consists of a larger and ever changing adaptive item bank, it can be individually administered, and it is interactive and time independent.

Nowadays, CAT is not only a field under research, but it is also used as an assessment tool for authorized foreign language examinations such as TOEFL. Though a promising field of study, little implementation in large-scale language assessment has taken place. This paper makes a review of CALT and reviews the implementations up to date. Then, it concentrates on the problem of the discrimination between the proficient and the good learner in multiple-choice (MC) questions. Finally, it presents a model for computer adaptive open-closed items, which will be able to discern proficient from good students.

COMPUTER ADAPTIVE TESTING (CAT)

CAT is a branch of Computer Based Testing (CBT) and Artificial Intelligence (AI) that provides personalized testing and more accurate results concerning the cognitive level of every individual. In other words, CAT is tailored to the ability and level of each examinee. Based on an algorithm, the computer can update the estimate of the examinee's ability after each item and select the next item on the basis of the new ability estimate. On the basis of the examinee's previous answer, the system acts as a human examiner and selects the next item which is of greater difficulty if the examinee has answered correctly, or of less difficulty in case the examinee gave a wrong answer.

Computerised assessment is wide spread nowadays due to the low cost of PCs and the fact that all public schools have networked computer labs, where pupils become computer literate from a young age. Finally, most international testing organizations have started delivering their tests in CBT and CAT mode worldwide via third-party trustees, making their tests available to even more people, more frequently and in less time.

CAT systems are student-centered as they – contrary to the P&P counterpart – can update the estimate of the examinee's ability, called User Profile, after each item and can be used in the selection of the subsequent items. They also have increased efficiency, greater precision with less items, longer duration as only a few items from the item bank are exposed. Thus, the tailored item selection can result in reduced standard errors and improved accuracy for scores for high and low ability test takers. Tailored item selection also leads to avoidance of examinee's boredom from answering too easy questions and of frustration from answering too hard questions. Problems associated with P&P tests like ambiguous answers or physical problems with the answer sheets (Wainer et al, 2000) are being solved. Moreover, these systems are time-effective, since fewer items are needed to achieve accuracy. CAT offers also greater test security (Wainer et al, 2000) than traditional P&P tests, as it is impossible for the examinee to know the items in advance. CAT shares all advantages of CBT, such as immediate feedback and self-pacing. Web-based CAT exploits the capacities of the net, such as on-line, immediate scoring, easily downloadable software, and item pool update.

Yet, research has revealed some drawbacks. Firstly, CAT, similarly to CBT, requires an equipped computer lab and computer literate examinees. Furthermore, CAT is not applicable to all subjects and skills, as it is based on the Item Response Theory model (IRT), which is not applicable to all item types. The fact that CAT requires careful item calibration renders it incapable of including items that cannot be easily calibrated, such as open-ended questions. Hardware limitations may also restrict the types of items that can be administered by the computer. Another drawback is that the examinees are not permitted to go back and change answers, as the program selects next items on the basis of the answered items, rendering reviewing implausible. Studies show that only when both P&P and CAT had the same test-taking flexibility (e.g. item review), test results were equivalent (Sawaki, 2001). CAT philosophy, however, prohibits reviewing. To sum up, CAT systems have both merits and flaws, and they cannot specialize on every plausible item.

CALT AND THE FOUR SKILLS

In foreign language testing, a well-constructed test needs to have some basic qualities. Test scores measure the degree of examinees' proficiency and should be valid and reliable. A test needs to have reliability in order to measure examinee's performance accurately and to certify the true ability of the examinee in various successive versions of the same test (Hughes, 1989). Any error of measurement in an individual's score is due to lack of reliability and in such a case the test is considered unreliable. Validity is the second test quality that shows whether the test has measured the intended skills and abilities it was constructed to measure (Bachman and Palmer, 1996). A test needs also to be practical and economic. Maximum quality with less effort and within less time is preferable. Economy in time and item selection can result in increased test production and higher scores from the part of examinees. Foreign language tests should measure evenly productive – speaking and writing – and receptive – reading and listening – skills and the subsets of every skill have to be analogous in length, level and time. Finally, test developers should consider the impact of their test, as it depicts the philosophy of the educational system, serves the needs of the society (Bachman, 1990) and can influence examinee's behavior.

Traditional CALT follows a patterned procedure. Test items are categorized in terms of levels of difficulty. The test starts with an item of average difficulty that corresponds to the level of the average student. If the item is answered correctly, the system selects an item of a higher level of difficulty, while in the opposite case, the chosen new item is less difficult than the previous. The test proceeds in the same pattern, until the stopping parameter occurs. The test score derives from the average level of difficulty of the items answered correctly. Item response times can also be monitored in order to be tailored to each examinee or give some information about the examinee's performance. Response times are different for each individual and reveal various traits and cognitive skills (Lamboudis and Economides, 2002, 2004). This supposition is against strict time limits, advocating that there should be given adequate time per item, in order for the test taker to decide calmly (Schnipke and Scrams, 1997). Research has also shown that response times for wrong answers are longer than those for correct answers (Hornke, 2000). Therefore, the more time an examinee spends on an item, the more prone he/she is to mistake.

Adaptivity on Vocabulary and Grammar Assessment

Vocabulary and grammar knowledge in FL are considered one of key factors indicating fluency. Different lexical and grammatical items are measured in various levels of language proficiency. Adaptivity of vocabulary items is therefore important and imperative. One way to assess vocabulary is by categorizing words into levels of competence. Some lexical groups presuppose the knowledge of others in lower levels, and this is an indication of the size of vocabulary each examinee has. Another way to assess vocabulary is by measuring the strength of vocabulary knowledge, which can be separated in four levels of difficulty, starting with the easiest: receptive recognition, receptive recall, productive recognition and productive recall. Studies in this field have shown that adaptive tests measuring vocabulary knowledge in terms of size and strength have managed to assess examinee's level of vocabulary knowledge accurately (Laufer, et al, 2001).

Adaptivity on Oral Production

Up-to-date, most large-scale administered adaptive tests do not have an adaptive component in oral proficiency. Yet, research in this field is taken and some systems are already in use. The Center for Applied Linguistics (CAL) has developed the Computerized Oral Proficiency Instrument (COPI) in Arabic, Chinese and Spanish, an adaptive test that gives many initiatives to the test-takers. Examinees are given more control of various aspects of testing and a self-assessment tool enables the system to extract more information about the examinee's oral proficiency. Thus, the CAT has plenty of information to rate and calibrate the examinee's performance, leading to scoring accuracy.

Adaptivity on Reading Passages and Listening Comprehension

CAT is widely used in large-scale examination for the assessment of examinees' FL reading and listening proficiency. Reading is a receptive skill and can be easily assessed in multiple-choice, close-item form or in the form of testlets, where one reading passage is followed by a group of questions, which have the same difficulty level. The important issues in the development of adaptive reading items regard the reading construct validity, the IRT theory used and the measurement of the items. Adaptive listening items assess examinees' ability to understand a range of oral speech, from short utterances, such as single words to short monologues and dialogues and to longer discussions (Dunkel, 1997).

Up-to-date Electronic Marking of Written Production

Open-ended questions and open writing tasks are still marked by human examiners, as there is no valid Natural Language Processing (NLP) technology to undertake electronic marking. However, new advances that have been made in text and speech recognition, enable electronic short answering, information retrieval and summarization with the use of semantic parsers, syntactic parsers, text mining, language databases and electronic corpuses (Harabagiu and Ciravegna, 2002). The Intelligent Essay Assessor (IEA) is a commercial grading software financed by the Army

Research Institute and developed at the Knowledge Analysis Technologies. Research has shown that it can assess essays as accurately as a human examiner (Streeter L. et al, 2002). Some large-scale tests, such as GMAT make partial use of e-marking programs, together with human readers. E-rater is am essay scoring e-marking program developed in ETS, and designed to mark the two types of essays of the GMAT examination holistically in a few seconds (Burstein, et al, 2003). The program marks together with a human reader, and it is regarded highly reliable (Burstein and Wolska, 2003). However, scores of essays written by non-native speakers of English had a slightly bigger variation between e-rater and human readers, showing that there are some non-native syntactic and semantic structures not evaluated by the e-rater (Burstein, and Chodorow, 1999). Though promising, e-rater cannot still be used exclusively, as studies have proven that such programs can be fooled by experts in writing (Powers, 2001). Future research on electronic paraphrasing and lexical metonymy may enhance its accuracy (Burstein, 1996).

CALT APPLICATIONS IN LARGE SCALE TESTING PROGRAMS

QPT (The Oxford University Press Quick Placement Test)

QPT is the official Oxford placement test that is issued in both P&P and CBT form. Based on ALTE's (The Association of Language Testers in Europe) standards whose main aim is to establish common levels of proficiency (the ALTE Framework) in order to promote the transnational recognition of certification in Europe (Jones, 2001), QPT is designed to calculate accurately English language learners' level of proficiency, from the beginner to the very advanced (Cambridge Proficiency Examination) stage. Its CBT version is adaptive, using item banking and IRT and it takes 15-20 minutes to administer, whereas its P&P counterpart takes 30 minutes to complete. Thus QPT CBT needs half of the time to make estimation. Another advantage over the P&P version is that QPT CBT can assess vocabulary, grammar and reading proficiency, and also listening comprehension. All items are in MC form, making the test easy even for computer illiterate examinees to complete. The existence of both versions can assert the reliability of the CBT version, as a failed examinee in the CBT version can reassess his/her level by taking the P&P version and compare the scores. Examiners can save time that can be dedicated to assessing examinee's oral and writing skills with face-to-face interviews and short essays. In such a way, examinees may have the opportunity to assess all language skills quickly and time-effectively.

TOEFL (Test of English as a Foreign Language)

The CBT version of the test was introduced in 1998 by the Educational Testing Service (ETS). TOEFL has a wide examinee population from divergent cultures. The adaptation of the system to this bulk of examinees is a great challenge, considering the cultural and linguistic diversity of the test takers. Today, the test is partly adaptive, incorporating a wide range of items and question types that can be found in its P&P version together with new items that exploit the visual and audio capabilities of ICTs. To prepare students for the CBT version, tutorial lessons have been developed, teaching examinees basic computer skills. Only the Listening and Structure sections of the test are adaptive, altering their level of difficulty by presenting error correction and MC items according to the examinees' performance. The reading section consists of an arbitrary and not adaptive selection of passages with questions. Finally, the writing section can be either computer-based or paper-based, marked by specially trained examiners via an on-line scoring network system (OSN) (Lee, 2001). Examinees have the opportunity to change answers as long as they do not confirm the final choice. When the confirmation button is selected, the examinee cannot go back to change the answer and the system selects the next item. The greatest challenge of TOEFL is how to guarantee validity and reliability in the unidimentional IRT calibration and scaling, and in the item selection algorithm for multicultural examinees.

CoRA (Contextualized Reading Assessment) - Minnesota Language Proficiency Assessments (MLPA)

The MPLA are second language assessment tools for reading, writing, listening, and speaking, available for French, German, and Spanish at two levels (Intermediate-Low and Intermediate-High level) on the scale developed by the American Council on Teaching of Foreign Languages (ACTFL), assessing minimal proficiency in a second language. The MPLA provide certification for entrance to, or exit from, a course of study. While the MLPA assess all four language skills and they are mainly combuter-based, as well as in P&P form, only the Contextualized Reading Assessment (CoRA) is adaptive. CoRA consists of a series of reading passages, each followed by a testlet, a group of items reffering to the passage. After the completion of each testlet, which has a specific level of difficulty, the IRT selects the nest testlet based on another passage. Testlets are scored dichotomously (0 or 1), using IRT. (Chalhoub-Deville, M. et al., 1996)

COPI (Computerized Oral Proficiency Instrument)

The Computerized Oral Proficiency Instrument (COPI) is a self-administering, adaptable, multimedia speaking test that allows examinees control over various aspects of the test situation. Its speaking tasks follow the model of the tape-based Simulated Oral Proficiency Interview (SOPI). However, unlike the SOPI, on the COPI examinees are given at least partial control over several aspects of the test administration. These aspects include amount of thinking and response time, speaking functions and topics to which to respond, level of difficulty of several of the tasks, and the language of task instructions. While examinees typically respond to seven tasks, a large underlying pool of tasks, together with the flexibility of multimedia, allows examinees this control. Speaking tasks on the COPI are grouped into four levels of difficulty. At the start of the COPI, examinees complete a self-assessment of their speaking ability. The outcome of the self-assessment provides both the examinees and the test administration software information regarding at which level of difficulty to start the test (Malabonga, 2000, Malabonga, and Kenyon, 1999).

BEST PLUS (Oral English Proficiency Test)

BEST Plus is the oral component of the Basic English Skills Test (BEST), developed by CAL. It has a computer-based, adaptive section and a semi-adaptive printed test booklet for the assessment of oral proficiency. In its computer-adaptive version, the algorithm selects the appropriate item out of a large pool, so that very few items can be repeated and the validity of the test is increased. Items are categorized according to personal, community and occupational language use domains, and the algorithm selects items related in terms of theme and level of difficulty in order to simulate the element of human communication (Stauffer and Kenyon, 2001). Answers are listened and scored by the test administrator, while the algorithm selects next items prompted by the previous score. The test can assess all proficiency levels in terms of listening comprehension, language complexity, and communication of meaning. BEST Plus was successfully pilot-tested and is widely used for multicultural learners of English.

GRE (Graduate Records Examination)

The Graduate Records Examination (GRE) is not a FL examination. However, it includes the Verbal Ability measure module, which is adaptive. In general, GRE General test aims to assess students' verbal, quantitative, and analytical writing skills in English, both as a native or foreign language. Two of the aforementioned skills – verbal and quantitative – are adaptively tested. The MC items of Verbal Measurement are separated in four sections: analogies, antonyms, sentence completions and reading comprehension, selected in arbitrary order. They measure the ability to use synthesis and analysis skills, understand terms and concepts, extract specific information and find relationships between words or sentences. Comparability studies in 1992 showed student acceptance of the CBT version and final equation of the CBT linear mode and its P&P counterpart (Schaeffer, et al, 1995). A second comparability study took place in 1998, that resulted in higher CAT mean scores for lower-scoring P&P groups, such as minority students and women, due to the implementation of the 80% scoring method (Schaeffer, et al, 1998), which led to the initiation of the proportional scoring method. The CBT version of Verbal Measurement is different from its P&P counterpart (GRE Practice General Test, 2004). The final score depends on the statistical characteristics of each item, the content covered, the variety of the items, and the item answered in the allotted time. Also, the CAT functions like a test assembler, managing also, among other things, item exposure and overlap, and conditional standard errors of measurement (CSEMs) (Schaeffer, et al, 1995). Scoring uses an IRT maximum likelihood theta estimation procedure, building the examinee's ability estimate after his/her performance. The final score is a scaled score that can be compared with the "number-right" true score (Schaeffer et al, 1995).

GMAT (Graduate Management Admission Test)

The Graduate Management Admission Test (GMAT) measures basic verbal, analytical writing and mathematical skills, acquired via school or academic education and work. The analytical writing section (AWA) is administered to assess analytic skills on processing issues and arguments. Two tasks are administered and should be analyzed in 60 minutes. The essays are scored either by two human examiners or by one human examiner and the E-rating electronic system by ETS. There is special treatment of non-native test-takers by the examiners. The verbal section of the GMAT has three types of MC items: reading comprehension, critical reasoning and sentence correction, assessing both cognitive and language skills. Reading comprehension assesses word understanding, logical relationship between concepts and arguments, inferences drawn, and understanding of qualitative concepts in verbal form. Critical reasoning assesses argument construction and evaluation and plans of action evaluation. Finally, sentence correction assesses correct and effective expression in lexical, grammatical, and structural terms. To ensure validity, GMAT scores are compared with students' grade point average (GPA) as well as other predictors (GMAC Validity Study Service 2002–03). Both GRE and GMAT developers acknowledge the fact that these tests cannot accurately measure the actual examinee performance on a working or academic environment, and make only estimations on students' abilities (Table 1).

a = adaptive n/a = not adaptive	**QPT**	**TOEFL**	**COPI**	**BEST PLUS**	**GRE**	**GMAT**	**MLPA**
Only for FLA	x	x	x	x			x
Reading	x (a)	x (n/a)			x (a)	x (a)	x (a)
Use of English	x (a)	x (a)			x (a)	x (a)	
Listening	x (a)	x (a)					x (n/a)
Speaking			x (a)	x (a)			x (n/a)
Writing		x			x (n/a)	x (n/a)	x (n/a)
Only MC	x	x			x	x	x
Certificate		x			x	x	x
Certificate in FL		x					x
P&P Version	x						x
Multimedia	x	x	x				
Online Demo		x			x	x	x
IRT	x	x			x	x	x
In Authoriszed Centers only		x			x	x	x
Levels	ALL	C1	ALL	ALL	(C2)	(C2)	B1, B2
Standards	ALTE	ETS	ACTFL	ACTFL	ETS	ETS	ACTFL

Table 1: Characteristics of the existing CAT systems in language assessment

THE PROBLEM

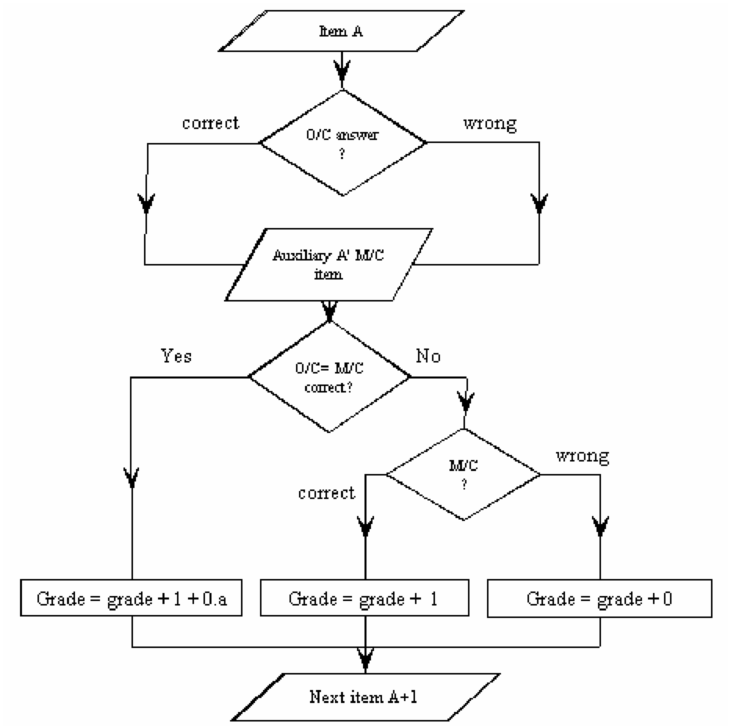
CALT technology can and should be student-centered. CALT nowadays is based on a solid programming that is collective rather than individualized and fails to include crucial cognitive parameters of student language competence and performance (Giouroglou and Economides, 2003). Such systems cannot replace the human examiner without nasty consequences for its group of examinees. We need to contruct systems what will not assess students horizontically as an equable lot but vertically as unequal individuals. Moreover, the new generation of assessment should create different experiences that will motivate test-takers, as it is proven that the new technologies are profoundly preferable to students, whenever they make the lesson interesting, motivating and interactive (Ali, 2001).

As described above, the vast majority of the current CALT systems use MC, close-ended items to discriminate among proficient, good and weak learners. This is mainly due to the fact that MC items are easily programmed and calibrated using adaptive technologies. 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. Examinees receive valid assessment, while the guessing parameter is decreased in proportion to the number of items presented. 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. This method is widely used by language testing organizations, such as the University of Michigan Certificates in English, while other organizations use a variety of MC and open-closed items, such as the Cambridge Syndicate and the State Examinations on Language Competence (KPG). Proficient learners answering MC items are not given the opportunity to discriminate themselves from good learners by openly typing the correct answer in an open-closed item in case they know it. They are forced to choose among the four intended choices and receive the same marks 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. The IRT functions equally for all students and the guessing parameter *c* applies even to those who do not guess. CAT programmers can easily prepare an additional alternative path for examinees, rewarding the proficient ones and not influencing the mark of those who fail to answer.

THE ADAPTIVE OPEN-CLOSED ITEMS METHOD

The method presented allows examinees to chose between two options. The introductory sentence or question of the item will be presented alone, without the multiple-choices. Two buttons below the item will allow examinees either to answer the question as an open-closed (O/C) item in the form of gap filling or constructed response (CR), or proceed to the multiple-choice (M/C) selection mode. If the examinee knows the answer and is able to produce it, then he/she will select the O/C mode and will have the opportunity to demonstrate his/her advanced knowledge. They can either type their answer in the O/C mode or choose the correct answer in the M/C mode. A correct answer will receive a bonus in the total score and will update the User Profile of the examinee. A wrong answer will not affect the final score and it will immediately direct the examinee to the M/C mode. When the M/C mode appears, the examinee will not be able go back to the O/C mode. The immediate selection of the M/C mode will not have a negative effect on the score, as correct choices will receive the highest mark (1), and the adaptive algorithm will immediately proceed to the next, increased difficulty item. Wrong choices will receive no mark and the next item will be easier. On the other hand, a correct answer in the O/C mode will give the student the highest mark and will reward him/her with a subdivision of the mark (0.a), depending on the level of difficulty of each item. This "bonus" or extra mark will counterbalance the effect of the guessing parameter *c* in the IRT. The item selection algorithm will proceed to the next item of increased difficulty.

Examinees that do not select the first option will not be negatively scored. This method will not affect the final score of the test or punish a wrong O/C answer (Flow Chart). Instead, it will give the opportunity to the examinees to demonstrate productive FL use and active FL extraction from their long-term memory. Moreover, the examinee will not know whether the item is answered correctly in the O/C mode. After filling-in the gap, the examinee will proceed to the M/C mode of the same item. This technique will ensure that in case the examinee gives a wrong answer, he/she will not be facilitated by the system by knowing that mistake. If the examinee knows that he/she has answered wrongly, and his wrong answer is also one of the distracters, then the program will have helped him and the validity of the answer will be decreased. All in all, this method aims to increase the ability discrimination among examinees, produce more valid ability estimation and decrease the guessing parameter that has a high percentage in MC items. This method is under development and will soon be pilot-tested to EFL examinees.



Flow Chart: O/C – M/C Item presentation algorithm

CONCLUSION AND FUTURE CHALLENGES

CALT research in the future will still be concerned with ways to administer valid and reliable tests that will assess in no time the four language skills. To this end, advances in a number of fields is required in order to administer tests able to assess both close-ended and open-ended items, as the validity of multiple-choice testing has been seriously criticized (Chapelle, 2001). Firstly, we need to fully explore and explain cognitive abilities regarding language learning. Cognitive Linguistics, Contrastive Analysis and Error Analysis studies will help CAT systems discern between examinees' errors and mistakes, knowledge and lack of knowledge, understanding and lack of understanding. Secondly, with the aid of AI, we will create assessment systems that will accurately measure competence in productive skills, such as speaking and writing. It has been acknowledged that up to date, no published speech recognition software can be adequately used for pedagogical or assessment purposes in language testing (Chapelle, 2001), as it does not possess basic communicative techniques and cannot understand the foreign speaker's interlanguage. Corpus Linguistics can also help towards this direction. Large language and interlanguage databases can instruct the computer on how to mark student responses more accurately. To sum up, in order to create valid CALTs, we need to adopt and work towards a new test theory (Mislevy, 1996), gathering information from various disciplines.

REFERENCES

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

Bachman, L. (1990), *Fundamental Considerations in Language Testing.* Oxford University Press.

Bachman, L. and Palmer, A. (1996), *Language Testing in Practice*. Oxford University Press.

Bailey, K. (1998), *Learning about Language Assessment. Dilemmas, Decisions, and Directions.* Newbury House Teacher Development.

Burstein, J., and Wolska, M. "Toward evaluation of writing style: Finding overly repetitive word use in student essays." In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistic.* April 2003, Budapest, Hungary.

Burstein, J., Marcu, D., and Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. In S. Harabagiu & F. Ciravegna (Eds.), *IEEE Intelligent Systems: Special Issue on Advances in Natural Language Processing, 18*(1), 32-39.

Burstein, J., Chodorow, M., and Leacock, C. "CriterionSM: Online essay evaluation: An application for automated evaluation of student essays." In *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*, Acapulco, Mexico, August 2003.

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.

Chalhoub-Deville, M. et al.. *An Operational Framework for Constructing a Computer-Adaptive Test of L2 Reading Ability: Theoretical and Practical Issues*. The Center for Advanced Research on Language Acquisition, University of Minnesota, July 1996.

Chapelle, C.A. (2001), *Computer Applications in Second Language Acquisition. Foundations for Teaching, Testing and Research.* Cambridge: CPU.

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

Giouroglou, H. and Economides, A. (2003), "Cognitive CAT in Foreign Language Assessment" *Eleventh International PEG Conference "Powerful ICT Tools for Learning and Teaching", PEG'2003*, 28 June-1 July, St. Petersburg, Russia.

GMAC Validity Study Service 2002–03, "Conducting a Validity Study" Graduate Management Admission Council, http://www.gmac.com/NR/rdonlyres/763E6D5D-6D94-4D1F-BADF-EB80B290AE4B/0/ConductingaValidityStudy.pdf

GRE Practice General Test, 2003/2004. Graduade Record Examinations Board. Educational Testing Service. ftp://ftp.ets.org/pub/gre/14614.pdf

Harabagiu, S. and Ciravegna, F. (Eds.), *IEEE Intelligent Systems: Special Issue on Advances in Natural Language Processing, 18*(1), 2002, pp. 32-39.

Hughes, A. (1989), *Testing for Language Teachers.* Cambridge University Press.

Jones, N. "Using Can-do Statements to Equate Computer-based Tests Across Languages." In the proceedings of the 23rd Annual Language Testing Research Colloquium, p. 30, held in St. Louis, Missuri on February 20-24, 2001.

Lamboudis, D. and Economides, A. "Adaptive Exploration of User Knowledge in Computer Based Testing", *WSEAS Transactions on Communications*, Vol.3, No.1, pp.322-327, 2004.

Lamboudis, D. and Economides, A. "Managing Time Thresholds in Mixed-initiative Learning Environments". In the proceedings *E-Learn 2002*, pp. 1746-1749, AACE 2002.

Linden, W. and Hambleton, R. (Eds.), (1996), *Handbook of Modern Item Response Theory.* Springer.

Malabonga, V. (2000). "Trends in foreign language assessment: The Computerized Oral Proficiency Instrument." *NCLRC Newsletter* (http://www.cal.org/nclrc)

Malabonga, V. and Kenyon, D. (1999). "Multimedia computer technology and performance-based language testing: A demonstration of the computerized oral proficiency instrument." In M. B. Olsen (Ed.), *Computer Mediated Language Assessment and Evaluation in Natural Language Processing.* New Brunswick, NJ: Association for Computational Linguistics.

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.

Sawaki, Y. (2001). *How examinees take conventional versus web-based Japanese reading tests.* Work in progress session presented at the 23rd Annual Language Testing Reserach Colloquium, March, 2001, St. Louis, MO.

Schaeffer, G.A. et al. "The Introduction and Comparability of the Computer Adaptive GRE General Test" *GRE Board Professional Report No.88-08aP*, Educational Testing Service, Princeton, New Jersey, August 1995.

Schaeffer, G.A. et al. "Comparability of Paper-and-Pencil and Computer Adaptive Test Scores on the GRE General Test" *GRE Board Professional Report No. 95-08P, ETS Research Report* 98-38, August 1998, Princeton: ETS.

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

Stauffer, S. and Kenyon, D. "A Computer-assisted, Computer-adaptive Oral Proficiency Assessment Instrument Prototype." In the proceedings of the 23rd Annual Language Testing Research Colloquium, p. 62, held in St. Louis, Missuri on February 20-24, 2001.

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.

Twigg, C. (1994), "The Need for a National Learning Infrastructure", Educom Review, 29.

Wainer, H et al. (2000) Computer Adaptive Testing. Second Edition. London: Lawrence Erlbaum Associates, Inc.

⇦