Giouroglou, H. & Economides, A. (2003). Cognitive CAT in foreign language assessment. In Proceedings 11th International PEG Conference, Powerful ICT Tools for Learning and Teaching, PEG '03, St. Petersburg, Russia, 28 June- 1 July, CD-ROM

Cognitive CAT in Foreign Language Assessment

Hara Giouroglou
Ph.D Candidate
hara@uom.gr

Anastasios Economides
Assistant Professor
economid@uom.gr

University of Macedonia,
Egnatia 156,
Thessaloniki 54006, GREECE
Tel: +30-31-891799, Fax: +30-31-891750

Abstract. Though revolutionizing, Computer Based Assessment (CBA) has not accomplished to become an established system of evaluation in Foreign Language Assessment (FLA). Research in CBA is very narrow and short-term, while the findings are usually misleading. This is mainly due to the inability from the part of New Technologies to simulate the human examiner and the lack of flexibility as regards language errors. In many cases, the performance of examinees during CBA varies from the corresponding performance during a traditional, paper-and-pencil examination. The reason for this has to do either with student or system inability. CBA systems are programmed to assess the competence of a wide number of individual students. Yet, our experience as teachers reveals that the examiner is not an impersonal "checker" but an active, emotional intervener between the examinee and the test. The paper aims to prove that Computer Adaptive Testing (CAT) can deservedly substitute the human examinee, by assessing the true cognitive state and foreign language competence of each student. CAT technologies can introduce a new, student-based era in Foreign Language Assessment that will be personalized, flexible, and sensitive to human cognition, language processing and error correction. To this end, research in FLA needs to fully exploit the current findings in CAT, Cognitive Science, Foreign Language Learning, and Error Correction. After reporting the findings in error correction, the paper separates wrong answers into errors and mistakes and proposes an innovative methodology for CAT as regards FLA that will approximate the exact language competence of each student individually.

Keywords: Computer Adaptive Testing, Foreign Language Assessment, Cognitive Science

1 Introduction

Over the last years and especially after the dissemination of the Internet for public use, a vast literature has been concerned with the effects of new technologies in Language Learning (LL) and remarkable research by authorities in the field of education has proven that the new instructional media have helped learners achieve better and teachers deliver more interesting, authentic and motivating lessons. Apart from the use of ICT during lesson delivery, new technologies have also penetrated the assessment phase. Computer Based Testing (CBT) is an official branch of CALL and research in this area aims to create systems that will measure language proficiency as accurately as traditional means of foreign language assessment (FLA). Close-ended questions can be easily authored and assessed in electronic form, while open-ended questions and compositions still need more specialized programming, while the accuracy of computer-based systems on specific language skills, such as reading comprehension, is stilled questioned (Chalhoub-Deville,1999). CBT, like CALL, should be researched interdisciplinary, adapting theories of Psychology, NLP, and Artificial Intelligence (AI), Human-Computer Interaction (ICT), Applied Linguistics, Cognitive Science (Levy, 1997).

Computer Adaptive Testing (CAT) is a branch of CBT and 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. The purpose of the article is to show the current imperfections of CAT for FLA and to propose new principles that will help such systems approximate the exact level of language competence of the examinees and "mimic automatically what a wise examiner would do" (Wainer, 2000), in order to achieve authenticity, construct validity, and measurement accuracy.

2 CAT Considerations in FLA

CAT systems are considered student centred as they – contrary to the paper-and –pencil counterpart – can update the estimate of the examinee's ability (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. Moreover, these systems are time-effective, since fewer items are needed to achieve accuracy. Finally, CATs share all the advantages of CBT, such as immediate feedback and self-pacing.

Yet, research has revealed some drawbacks. Firstly, CAT, similarly to CBT, requires an equipped computer lab and computer literate examinees. Furthermore, CATs are not applicable to all subjects and skills, as they are based on the Item Response Theory model (IRT), which is not applicable to all item types. IRT is a mathematical function of the examinee's proficiency parameter (θ) and three item parameters (a = item discrimination, b = item difficulty, c = guessing parameter), predicting the probability of the examinee's success. Then, the item-choice algorithm selects the item on behalf of the examinee's earlier answer. We conclude that IRT may accurately operate in close-ended items, such as multiple-choice questions. Yet, we cannot assume that the four IRT parameters can accurately estimate the examinee's competence on every skill of a FL.

The fact that CATs require careful item calibration renders that incapable of including items that cannot be easily calibrated, such as open-ended questions. Apart from that, hardware limitations may restrict the types of items that can be administered by the computer. Another crucial 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. 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, and in many cases examinees who sat both FL paper-and-pencil and CATs failed or achieved low marks in computerized testing. To sum up, CAT systems have both merits and flaws, and therefore they cannot specialize on every plausible item.

3 The Human Examiner in FLA

During traditional education, the teacher of the student-centered language classroom adapts himself/herself to the needs, knowledge and learning style not only of the class as a whole, but also of each student separately. The dynamics of each class are divergent, namely because they consist of a different set of students. Apart from that, each student has his/her own personality, experiences, as well as cognitive style. Therefore, not all students should be approached equally, but they should be treated individually and not collectively. To conclude, the objective teacher of the modern era follows a flexible approach, focusing on the actual knowledge of each student, rather that sticking on minor, trivial details that might impede students' will to learn.

Likewise, the human examiner should also follow the same student-centered approach that will bring to the surface the true foreign language ability of each student. The adaptation of the examiners is a common secret among educators, who tend to be facilitators rather than authorities. In order for the examiner to adapt to each examinee's learning capacity, he/she should be aware of the following variables: student's age, language learning background, learning style, native and foreign language performance, and *whole-test performance,* which is the ability of the examiner to process not only each test item separately but also all test items as a whole, in order to mark errors done not due to ignorance but due to time constraints, hurry, negligence, or even cognitive inability (e.g. dyslexia). Thus, marking becomes more objective and student friendly, promoting learner motivation while eliminating learner discouragement.

4 The Problem

By replacing CBT with CAT in FLA, we can move one step forward and create student-centered systems that can simulate up to some point the human examiner, by creating User Profiles and selecting the right items. This is not enough, however. The human examiner is not only an intervener between the examiner and the test, but also an objective error corrector. He can cognitively evaluate student responses and detect whether an error is due to ignorance or negligence. This knowledge enables the examiner to mark more objectively, approaching more accurately the foreign language performance of the examinee.

The proposed CAT system for FLA is a testing program that will differentiate between errors – made due to ignorance – and mistakes – made due to negligence. Thus, the system, apart from correcting the traditional multiple-choice items, will be able to process accurately more open items, such as gap filling, open-ended questions of even essays. The idea is to create a system that will not be based on a behavioristic, "yes or no", "0 or 1", "true or false" model, as it does not correspond to the human brain activity during language performance. We are inspired to develop cognitive CAT systems for FLA that will be adapted to each examinee's true ability based on a pre-test questionnaire, and performance based on the examinees actual behavior during the test.

5 Findings from Psycholinguistics

Studying the findings of psycholinguistics regarding speech production and comprehension we will analyze the construction of more accurate CAT systems that will try to measure the exact linguistic level of the examinee.

Psycholinguistics examine the inner processes of the human mind that lead to linguistic proficiency and language acquisition. There is a vivid relationship between language, thought and cognition as Chomsky and Piaget have advocated from different points of view (Chomsky, 1972, Rieber, R. W., & Voyat, 1983). Historical brain research is proven to be directly combined with first and foreign LL (LeLoup & Ponterio, 2003). Gardner's research on Multiple Intelligences also proved individual inclinations in learning (Gardner, 1983). Research in first language acquisition has revealed serious findings in human language development that can also be applied in second/foreign language acquisition.

Linguists divide wrong answers in errors, which are systematic, and mistakes, which are non-systematic (Richards, 1993). The systematic errors together with the actual speech production that involves speaking and writing (productive/expressive skills) can be examined by the analysis of hesitations, speech errors (slips of the tongue, slips of the pen, slips of the hand in Sign Language) and language disorders (Akmajian et al. 1998). Boomer and Laver define a slip of the tongue as "an involuntary deviation in performance from the speaker's current phonological, grammatical or lexical intention" (Ellis & Beatie, 1996). Hesitations and speech errors are natural phenomena as there is always a conflict between the structured, serial nature of a language and the abstract, untamed nature of thought. Hesitations regard the time span between the input and the output. In oral speech, learners may stay silent (silent pausing), produce meaningless utterances (um.., er..), or lengthen the sounds for some time and in written speech, they may take time to write a sentence or even a word. This reaction may be either spontaneous (errors), if this is the way they generally behave even in their mother tongue, or intended – usually in prolonged hesitations –, if they have gaps in language competence (mistakes). However, this does not happen solely during speech, as "it is as true of speech as of any other sphere of cognition" (Ellis & Beatie, 1996).

If hesitations and speech errors are common facts in first language articulation, they should be even more frequent in foreign language articulation and production, either written or spoken (Richards, 1993). Foreign language production is a more time-consuming mental process, as the desired output needs to be consciously controlled before being actually produced. In a foreign language learning environment speech errors can be either spontaneous or induced according to the level of each learner's proficiency. In the first case, they are simple errors due to anxiety, time pressure or confusion, while in the second case, they are mistakes and they are done repeatedly due to improper language use or learning. The most frequent speech errors involve linguistic constituents and include:

a. *Exchange errors* – foon speeding
b. *Anticipation errors* – a cuff of coffee
c. *Preservation errors* – John gave the goy
d. *Blends* – omnipiscient
e. *Shifts* – Even the best team losts (teams lose)
f. *Substitutions* – "confession" for "convention" (form), "yesterday" for "tomorrow" (thematic unit), "finger" for "toe" (meaning).

It is obvious that these errors are rather anticipated and not random in their nature. Therefore, we can assume that learners making such errors are not ignorant of the syntax, morphology, semantics and pragmatics of the target language but are usually confused, either due to a natural inclination (they also make slip of the tongue in their mother tongue.) or due to their level of proficiency.

Apart from hesitation and speech errors, there is apparent confusion between first and foreign language production. In interlingual errors (Larsen-Freeman & Long, 1997), the dominant language capacity of each FL learner can be an obstacle to fluent FL production. Phonetic, lexical, semantic, pragmatic, syntactic or grammatical mistakes in FL may be due to first language influence. For example, the fact that the word “dramatic” means “tragic” in Greek and “theatrical” in English causes confusion to Greek learners of English. The FL errors associated to the first language of the learner can be identified and anticipated according to Contrastive Analysis research (James, 1997).

Another source of language errors stems from the mental lexicon. There words are categorized in terms of phonetic, semantic or orthographic resemblance. Common mistakes are: “witch” instead of “which”, and “whith” instead of “with”. In the case of a foreign language, its mental lexicon is highly influenced by the first language. Greek, for example, is a language that orthography matches pronunciation. Thus, many Greek students have difficulty in understanding the phonetic and orthographic correlations of the English language.

6 The Study

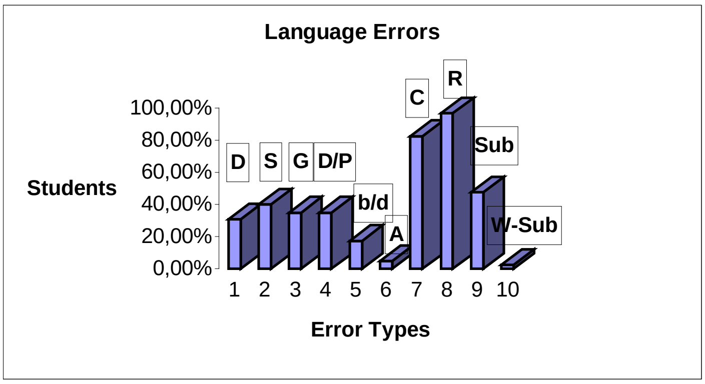
Up to this point, we separated language errors into “mistakes”, done due to ignorance, and “errors”, done due to negligence. We also categorized errors into cognitive misinterpretations of the final outcome. Finally, we moved one step forward to assume that such errors can happen in all kinds of FL tests, even computer-based. In the case of CAT though, such a circumstance would have nasty results as the system would immediately assume “examinee ignorance”, lowering the level of item difficulty and giving a false score.

In order to present actual data on first and foreign language errors, we constructed a questionnaire that was distributed to 120 Greek high-school students of the Hellenic College of Thessaloniki, aged 12-17 years old. The objectives were firstly to detect the most frequent learning styles, secondly to locate their main FL errors and performance during a FL test and thirdly to figure out whether they make similar errors in their mother tongue. The vast majority of students were right-handed, and only 11 students were left-handed. Yet, there was no particular answer divergence between left-handedness and right-handedness. No student suffered from dyslexia.

We classified learning styles in 8 categories with reference to Gardner’s findings on Multiple Intelligences (Gardner, 1993, Amstrong, 2000). The results showed that the majority of teenage students are Kinesthetic (19%), Inter-Personal (18%), Spatial (15%), Musical and Intra-Personal (14%). 10% of the students being surveyed were Logical-Mathematical, while only 4% of them showed Linguistic Intelligence. These findings show the relevantly low linguistic capacity of teenagers in comparison to the other intelligence types.

Proceeding to the main questionnaire outcomes [Figure 1], we categorized some of the most frequent language errors we have come across as examiners. Dictation anticipation errors [e.g. *tommorow*] were reported to 30,8% of the students. 40% of the students admitted doing interlingual syntactic errors [e.g. *I tomorrow will go out*], while 35% were prone to grammar errors [e.g. *I am usually playing tennis*], Figure 1.

Figure 1



The next set of questions dealt with cognitive errors, usually done due to negligence or cognitive overload. 35% of Greek students confuse English dictation with English pronunciation. Accordingly, the examiner can detect mistakes of the following nature: merige [marriage], hauz [haus], filosofy [philosophy]. Apart from that, 17,5% of the students admitted confusing the letter *b* with the letter *d*. This is a very crucial issue that human examiners usually overcome without negative consequences for the examinees. Anagram mistakes [e.g. *nad* for *and*] were reported to only 5% of the students. Moreover, almost half of the students [47,5%] reported doing frequent substitutions – “slips of the pen” – such as *witch* [which] and *cut* [cat] due to orthographic or phonetic resemblance. These students are also inclined to produce inexistent wh-words such as *whith* [with], because of close orthographic resemblance to the existent wh-words. The fewest mistakes [2,5%] were reported in word group substitution, e.g. *apple* for orange. On the other hand, almost all students answered positively in two very crucial questions. 82,5% of the students spot and correct negligence mistakes when they make a final test scanning and 96,6% may remember a previously accessed item later. When asked if they also make the above mistakes in Greek, 24% of the students answered positively.

7 The Application of the Findings in CAT

The above findings suggest that a FL CAT test would not accomplish to detect the actual linguistic competence of a high examinee percentage. Students with Linguistic Intelligence would get more objective scores, but they only constitute a student minority. Apart from that, a respectable number of students would be prone to "slips of the key" and cognitive errors, due to negligence, cognitive overload, wrong interpretation of the item question, tiredness or nasty environmental conditions, without having the opportunity to correct later. Finally, the vast student majority would be deprived of the privilege to reassess their answers and substitute the previously wrongly accessed item with the correct one.

We should not disregard the fact that almost the one-fourth of the students questioned admitted doing similar mistakes or having similar performance in their first language. This group of students would get the most unjust score both in first and second language CAT. However, this does not mean that these students do not know their mother tongue; it implies that for various reasons they have difficulty in accessing linguistic items accurately. Even in the case of a mistake, language testers need to agree that the utterance or string of writing is not completely wrong, and mark it accordingly. In pen-and-paper tests, teachers usually pay no or little attention to such errors as they are cognitively competent to understand the intended meaning. However, computers are very strict markers as they cannot judge cognitively.

Consequently, as designed and operating today, CAT systems in FLA cannot assess the examinee linguistic competence in FL but the examinee Linguistic Intelligence. This means that the examinees that have difficulty or need time to access a particular language item from their memory or their mental lexicon, would probably fail a CAT. Therefore, researchers need to review these findings and develop CAT systems that will measure language performance instead of student cognitive-mental ability.

8 A Methodology for FL CAT Systems Development

CAT technology, as being evolved today, is not student-centered. It 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. CATs nowadays are universal and have a wide number of target examinees. Such systems cannot replace the human examiner without nasty consequences for its group of examinees. Therefore, CATs as administered nowadays not only are not entirely "adapted" to examinees, but may also produce false results, because they ignore the most important cognitive abilities of the human mind. To this end, we need to revise and add to the principles governing CAT in FLA.

The anticipated nature of these errors makes their programming on computers feasible. With the aid of adaptive technologies we could create CAT systems that will be able to identify "slips of the pen" or more specifically "slips of the key" or "slips of the click" and mark them accordingly. Developing such systems we may succeed in making CBT more objective and human-like in foreign language assessment.

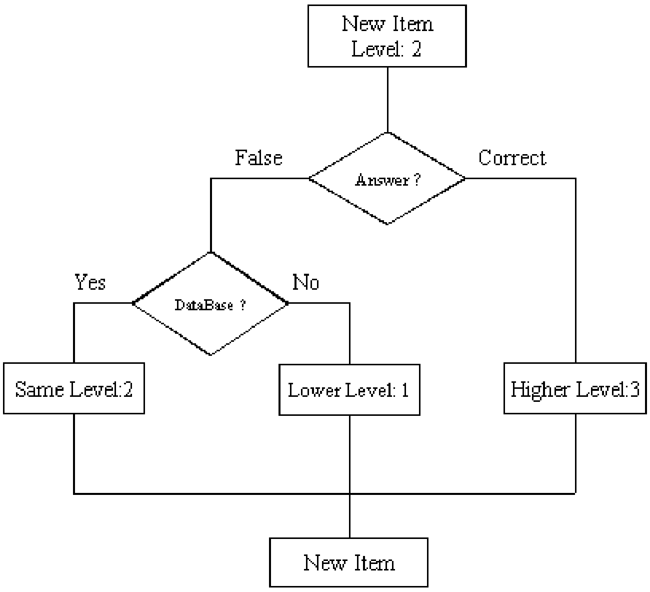
A pre-test questionnaire on learning styles and both first and foreign language performance can construct a preliminary User Profile that will help the system anticipate student behavior. For example, examinees with low linguistic intelligence will be given more time on every item, or will be automatically corrected when they make minor errors (e.g. *b* for *d*). On this basis, the Language-CAT (L-CAT) will show greater tolerance to students who are prone to language errors, with the use of an algorithm that will be able to discern errors from mistakes.

Traditional CATs follow 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 comes. The test score derives from the average level of difficulty of the items answered correctly. The proposed L-CAT goes one step forward this bilateral procedure.

In order for the cognitive L-CAT to separate between common errors and mistakes, the programmer should construct a database of the common errors of a specific target language that will be adjusted on the L-CAT. These errors are both the language-specific "slips of the key" as listed above and the interlingual errors. Therefore, in order for an L-CAT to have accurate results, it needs to be tailored to a specific target group of examinees, e.g. Greek learners (a wrong answer may be regarded as an error for a

Greek examinee – due to interlingual interference – or as a mistake for a Chinese examinee). This principle is also applicable to paper-and-pencil tests, as "it is questionable that FL tests should be and need to be universal" (James, 1997). If this is true for traditional assessment, it is twice as important for L-CAT not to be universal but adaptable to the first and foreign language of the examinees.

Figure 2



During the test, the L-CAT will firstly evaluate students' answers as correct or false [Figure 2]. The correct answer will follow the same route as in traditional CATs. False answers will be scanned by the common errors database. If the given answer matches a database systematic error, it will be regarded "correct" and the system will proceed with the next item selection without lowering the level of difficulty. Consequently, the proposed L-CAT will have three options: to increase the level of difficulty after a correct answer; to decrease the level of difficulty after a wrong answer; or to maintain the level of difficulty after an answer with a systematic error. Therefore, the overall score will not be influenced by the common language errors examinees usually do when they sit exams, and fluent examinees will still get distinctions.

9 Conclusion

In the era of communication, being able to interact in a foreign language is a more tangible and utile objective than spending years for FL mastering. The EU promotes plurilinguilism, a new term that presupposes the development of cognitive skills that will help learners communicate effectively in foreign languages or understand foreign words by making analogies from the bank of universal linguistic principles.

Assessment has immediate outcomes on LL. When learners are strictly marked, they may change their attitude towards the target language, may feel incompetent to achieve a high mark, loose motivation and self-esteem, and denounce the target language. In the threshold of the new era in education, cognitive learning needs to be followed by cognitive assessment methods. If we accomplish to locate individual cognitive competencies or impotencies in language reception and production, we will be able to focus on each student individually rather than on errors collectively. Therefore, language assessment will be a dynamic and not a static, passive or mechanic process that is not humanlike and therefore cannot produce accurate results.

Acknowledgements

Special thanks to Mrs. Ioanna Founta, the head of the Hellenic College of Thessaloniki, for her willingness to contribute to this research.

References

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

Armstrong, T. (2000). *Multiple Intelligences in the Classroom*.2nd ed. Alexandria, VA: Association for Supervision and Curriculum and Development.

Chalhoub-Deville, M. (Ed.), (1999), *Issues in Computer-Adaptive Testing of Reading Proficiency*. Cambridge: Cambridge University Press.

Chomsky, N. (1972). *Language and Mind.* New York: Harcourt Brace Javanovich.

Ellis, A., Beattie, G. (1996) *The Psychology of Language and Communication*. Fifth Edition. Exeter: Psychology Press.

Gardner, H. (1983). *Frames of mind: The Theory of Multiple Intelligence*s. New York, NY: Basic Books.

Gardner, H. (1993). *Multiple Intelligences: The Theory in Practice*. New York, NY: Basic Books.

James, C. (1997) *Contrastive Analysis*. Fourteen Impression. Essex: Longman.

Larsen-Freeman, D., Long, M.,H., (1997) *An Introduction to Second Language Acquisition Research*. Nineth Impression. Essex: Longman.

LeLoup, J.W., Ponterio, R. "ON THE NET. Foreign Language Study and the Brain" in *Language Learning & Technology,* January 2003, Volume 7, Number 1, pp.2-3.

Levy, M. (1997). *Computer Assisted Language Learning Context and Conceptualization*. Oxford: Clarendon Press.

Richards, J.C. (1993) *Error Analysis*. Twelfth impression. Essex:Longman.

Rieber, R. W., & Voyat, G. (Eds.), (1983) *Dialogues on the Psychology of Language and Thought: Conversations with Noam Chomsky, Charles Osgood, Jean Piaget, Ulric Neisser, & Marcel Kinsbourne.* New York, NY: Plenum Press.

Sawaki, Y. "Comparability of Conventional and Computerized Tests of Reading in a Second Language", in *Language Learning & Technology,* May 2001, Vol. 5, Num. 2, pp. 38-59.

Wainer, H. (Ed.), (2000), Computer Adaptive Testing. A Primer. Second Edition. Lawrence Erlbaum Associates, Inc.