Self-Assessment in a Feasible, Adaptive Web-Based Testing System

Eduardo Guzmán and Ricardo Conejo

***Abstract*—Adaptive testing systems generate tests for assessment that are tailored to each student. In these tests, students are assessed through a process that uses Item Response Theory (IRT), a well-founded psychometric theory. This theory is responsible for estimating student knowledge, determining the next question that must be posed at each moment, and deciding test finalization. System of Intelligent Evaluation Using Tests for Teleeducation (SIETTE) is a Web-based environment for generating and constructing adaptive tests. In SIETTE, teachers can create tests for self-assessment. In this kind of test, questions are posed one by one, and the correction of each question is shown immediately after the student's answer. Along with this correction, and in terms of the student's answer, feedback is provided. Feedback consists of pieces of knowledge that help students detect misconceptions or reinforce concepts correctly learned. Furthermore, hints can be included when questions are posed to supply students with some kind of help or explanation about the stem. As a result, this kind of test can be used not just for assessment, but also for instructional purposes. The first goal of this paper is to show how SIETTE can be used for instructional purposes, by combining adaptive student self-assessment test questions with feedback and hints. This paper also shows that the Web is a feasible platform for the generation of adaptive tests, supporting the use of SIETTE for this purpose.**

***Index Terms*—Adaptive testing, intelligent tutoring systems, Item Response Theory (IRT), student knowledge diagnosis, Web-based learning systems.**

I. Introduction

ANY instructional process should be complemented with assessments of the degree of assimilation of the topics (or concepts) studied. In intelligent tutoring systems, assessment is even more relevant, since these systems require some knowledge information source to guide the instruction. Ideally, each exam should be adapted to the personal circumstances of student. In principle, from a practical point of view, adaptation can only be carried out in environments in which the number of students is very small.

Computerized adaptive tests (CATs) represent an attempt to automate this difficult task, since with them students can be assessed independently. The main idea of a CAT is to act the same way a teacher would [1] when he or she assesses orally. If a teacher asks a student a question (so-called *item*) that turns out to be too difficult for him or her, the next question to be posed must be easier, and vice versa. One of the main disadvantages of CATs is that items are modeled with probabilistic functions. The values of these functions are inferred from performances of students that have taken this test nonadaptively. Accordingly, in order to be reliable, a CAT generation system should be able to collect valid student performance information to accomplish this inference.

System of Intelligent Evaluation Using Tests for Teleeducation (SIETTE) is a Web-based assessment environment that arises from an attempt to merge CATs and Web technology. This system can be accessed and tested at http://www.lcc.uma.es/SIETTE. SIETTE generates and constructs CATs for students. Once a student has finished a test, SIETTE returns an estimation of the student's knowledge level for each topic involved in the test.

Test-based assessment is essential to achieve an optimal learning process [2]. Although, in principle, tests are only used for assessment, they can also be used for instructional purposes because of self-assessment tests. In these tests, students are also asked items. The main difference with other tests is that, following the student's response, the item correction is shown within some feedback. This feedback provides the student with the reasons why his or her answer is correct or incorrect. In the case of an incorrect answer, a guide is supplied to identify the correct answer. Therefore, by using this kind of test, students are involved in a Socratic learning process [3] in which they take an active part in their own learning instead of merely receiving instruction passively.

This paper shows how SIETTE can be used to administer self-assessment CATs. An overview of the SIETTE system is shown, focusing especially on how the virtual student classroom generates adaptive tests for self-assessment. In addition, and to guarantee that SIETTE is a reliable CAT generation system, an empirical study was conducted. This study compared item calibration using data collected through SIETTE with the calibration carried out using data collected through paper-and-pencil (P&P) media.

In Section II, the theoretical fundamentals of SIETTE are presented in the CAT and the Item Response Theory (IRT). In Section III, the architecture of the system is briefly depicted. This section also includes a description about how CATs are generated in SIETTE. Section IV is devoted to the subject of language processors. Section V shows the comparative calibration study to evaluate the reliability of SIETTE. Section VI collects some other Web-based testing systems currently available, focusing on a brief description of their features. Finally, Section VII sets forth the conclusions obtained from this work and the research lines currently being followed.

Manuscript received August 2, 2004; revised June 20, 2005.

The authors are with the Escuela Técnica Superior de Ingenieros (ETSI) Informática, Universidad de Málaga, 29071, Málaga, Spain.

Digital Object Identifier 10.1109/TE.2005.854571

II. Theoretical Basis

A. Adaptive Testing

A CAT is a measurement tool administered to students by means of a computer instead of the conventional P&P format. In CATs [4], the presentation of each item and the decision to finish the test are dynamically adopted, based on students' answers. In more precise terms, a CAT is an iterative algorithm where items are posed one by one. This algorithm starts with an initial estimation of the student's knowledge level and consists of the following steps:

1) All the items in the knowledge base (that have not been administered yet) are examined to determine which is the best item to ask next according to the current estimation of the student's knowledge level.
2) The item is asked, and the student responds.
3) According to the student's answer, a new estimation of his or her knowledge level is computed.
4) Steps 1)–3) are repeated until the stopping criterion defined is met.

CATs select the next item to be posed, depending on the estimated knowledge level of the student (obtained from the answers to items previously administered). Selecting the best item to ask (given the knowledge level estimated) can improve accuracy and reduce test length. Different criteria can be used to decide when the test should finish, depending on the purpose of the test. A CAT can finish when a specified target measurement has been achieved, when a fixed number of items have been presented, when the time has finished, etc. Both the item selection and test finalization criteria are based on well-founded procedures that can be controlled with parameters that define the required accuracy. Furthermore, each student usually takes different sequences of items, different items, and even a different number of items.

The set of advantages provided by CATs over traditional P&P tests is addressed in the literature [4]. The main advantage of CAT is that it reduces the number of questions needed to estimate the knowledge level of the student, and as a result, the time spent on it. Furthermore, an improvement in student motivation is obtained. The estimation accuracy is much higher than the estimation achieved by randomly picking the same number of questions [5]. Tests are tailored to the particular features of each student. Computerized administration allows the inclusion in tests of a great number of item formats, taking advantage of multimedia facilities, such as sound, video, and high-quality images. In addition, item selection, test finalization criteria, and the estimation of the student's knowledge level can be performed efficiently and faster. Consequently, when students are offered the choice between a P&P and a CAT version of the same test, typically they prefer the latter option [5].

However, CATs entail some disadvantages, such as security. Examinees can memorize test items and share them with other future students. To overcome this problem, huge item pools are needed, along with techniques to control item exposure and to detect compromised items.

In CATs, the response model plays the most important role. This model describes how students answer the items depending on their knowledge level. Accordingly, this theory is used to estimate the student knowledge level from the response to each item. In addition, the response model can be used to determine the next item to be posed and to decide when the test must finish. In general, adaptive tests use IRT as a response model.

B. Item Response Theory

IRT has become the main basis of this measurement theory. IRT [6] rests on two principles [7]: 1) the performance of a student in a test can be explained by a set of factors called *latent traits*, which can be measured by means of unknown fixed numerical values and 2) the relationship between student item performance and the set of underlying item performances can be described by a monotonically increasing function called the *item characteristic curve* (ICC). The ICC represents the conditional probabilities of the successful answer to the item by a student with a certain *latent trait* (θ) measured in the domain of real numbers. The ICC must be previously known for each item and is expressed by means of a probabilistic function. In the field of CATs, the latent trait is the knowledge level.

There are several IRT model classification criteria, for example, in accordance with the number of latent abilities simultaneously measured, depending on the shape of the ICC, etc. As a result, there are many IRT models. The model implemented in SIETTE is one of the most commonly used, namely the three parameters logistic model (3PL) [8]. The ICC in the 3PL is modeled according to the following equation:

$$P_i(\theta) = P(u_i = 1|\theta) = c_i + (1 - c_i)\frac{1}{1 + \mathrm{e}^{-1.7a_i(\theta - b_i)}}. \quad (1)$$

$u_i = 1$ indicates that the student has answered item i correctly. In another case $(u_i = 0)$, the probability is equal to $1 - P_i(\theta)$. This model receives its name because it is characterized by the three parameters below.

1) *Discrimination factor* (a_i): A high value indicates that the probability of success by students with greater knowledge than the item's difficulty is higher.
2) *Difficulty* (b_i): This parameter corresponds to the knowledge level in which the probability of answering correctly is the same as answering incorrectly.
3) *Guessing factor* (c_i): This probability suggests that a student without any knowledge will answer the item correctly and represents the case in which a student answers randomly.

IRT has been successfully applied to the item selection mechanisms and the student's knowledge level estimation in CATs. The results obtained are independent of the tool used. This measurement is invariant with regard to the sort of test and to the individual who takes the test.

The main drawback of IRT is that the parameters of the ICC must be previously known for each item. These parameters are obtained initially from estimation techniques. The problem is that these techniques use the performances of students that have taken a test with these items nonadaptively.

III. SIETTE

As mentioned previously, SIETTE is a Web-based assessment tool. By means of a Web browser, teachers can create and

modify tests and items, and students can assess their knowledge by taking these tests. SIETTE generates CATs; therefore, all the steps of the life cycle of an adaptive test have been implemented.

A. Architecture

The architecture of the system [9] comprises the main components of CAT-based systems. The following parts can be mentioned.

- *Knowledge base*: Its contents are organized in subjects (or courses). A subject is broken down hierarchically (in a tree) in topics (or concepts), forming a curriculum. Items can be defined or associated to the topics. SIETTE can include different types of items [10]: true/false, multiple choice, multiple response, fill-in-the-blank, etc. Test specifications are defined in accordance with the topics they assess.
- *The student model repository*: This repository is a collection of student models. Each student model stores the information about a student's test session (knowledge level estimation, item posed, exposure time per item, etc.). The test generator dynamically updates these data after each response. The item calibration process can be carried out from the information stored in the student models.
- *Student classroom*: This virtual environment is where students take tests.
- *Test generator*: This component dynamically constructs test sessions. A test session is a test tailor-made for a student according to the specifications stored in the knowledge base.
- *External connections interface*: SIETTE not only works as an independent assessment tool, but also can be integrated into other Web-based tutoring architectures as an additional diagnosis module, providing well-founded assessments.
- *The authoring tool*: This Web-based utility adds and updates the contents of the knowledge base.
- *Result analyzer*: This utility allows teachers to analyze the performance of students in tests. Moreover, this tool provides a tool to consult the student models.
- *Item calibration tool*: This module uses the information obtained from the student model repository to calculate the ICC parameters.

Two different user profiles can take advantage of SIETTE. On the one hand, teachers can define subjects, structured in topics with items, and specify tests, making them available to students. They can also use SIETTE to make academic evaluations. In a controlled environment (such as a laboratory with PCs connected to the Internet), students can be assessed with SIETTE. The use of SIETTE has some advantages: tests tailored to students' personal features, automatic correction of tests, online grading generation, few software requirements (only a Web browser tool is needed), etc. To prevent cheating and unauthorized accesses, SIETTE incorporates several security mechanisms, such as test access restrictions by groups, Internet protocol (IP) addresses, or users. Finally, they can analyze the performance of the students that have taken tests. On the other hand, students may use SIETTE as a way to self-assess their ability in the subjects taught by teachers.

B. CAT Administering

When a student begins a test, his or her student model is retrieved from the student model repository. If there is no previous information stored about him or her (from earlier test sessions), his or her student model is initialized as a constant knowledge probability distribution curve.

During a test session, the next item to be asked of the student is selected adaptively by one of the following alternatives (decided by the teacher when preparing the test): 1) a *Bayesian criterion*—starting from the distribution of the estimated student knowledge, the selected item is the one that minimizes the sum of the *a posteriori* variances resulting from a correct or incorrect answer to the item or 2) a *difficulty-based criterion*—this method selects the item whose difficulty is closer to the estimated knowledge level of the student.

In general, in CAT systems, when a test of multiple topics is administered, the teacher must indicate manually the percentage of items that must be posed from each topic. This manual operation is unnecessary in SIETTE because the adaptive item selection engine is able to make a topic-balanced selection by itself. The selection procedure is accomplished in two steps [11]. First, the topic for which the estimate of the student's knowledge is lowest is selected. In the second stage, the most informative item from this topic is selected.

Once an item has been selected, it is removed from the set of available items for this test session. Next, the item is converted into a Web page and shown to the student. Optionally, if included by the teacher in this item, the student may be provided with a hint. The goal of hints is to supply students with some kind of help or explanation about the stem that will permit them to answer the item correctly.

After each answer, the student's knowledge level is computed using a Bayesian method [12]. In this method, the *a posteriori* probability distribution of the student's knowledge $(\mathrm{P}(\theta|\mathbf{u}))$ is calculated by Bayes' rule. Thus, the estimation process can be approximated to the following product:

$$\mathrm{P}(\theta|\mathbf{u}) = \prod_{i=1}^{n} P_i(\theta)^{u_i} \left(1 - P_i(\theta)\right)^{(1-u_i)} P(\theta|u_1, \ldots u_{i-1}) \quad (2)$$

where $P(\theta|u_1, \ldots, u_{i-1})$ represents the student's knowledge distribution inferred from the items previously answered in the test, and $\mathbf{u} = u_1, u_2, \ldots, u_i$ is the pattern of an examinee's responses in the i items of the test. Equation (2) gives a probability distribution curve as a result. The student knowledge level is equal to the mode of this distribution.

The teacher sets the finalization criterion during the test creation stage. In the test preparation phase, he or she must first indicate a minimum and a maximum number of items. When this maximum number is reached, the test finishes, and the estimation at that point of a student's knowledge level becomes the final estimation. In addition, SIETTE offers the following adaptive finalization criteria: 1) the most likely value of the estimated knowledge distribution is bigger than a certain threshold

1. Introduction
2. Lexical Analysis
   - 2.1. Lexical analyzer functionality
   - 2.2. Lexemes, regular expressions and tokens
   - 2.3. LEX
3. Syntactical Analysis
   - 3.1. Syntactical analyzer functionality
   - 3.2. Syntactical Analysis types
   - 3.3. HEAD and FOLLOW functions
   - 3.4. LL(1) analysis
   - 3.5. SLR(1) analysis
   - 3.6. LR (1) analysis
   - 3.7. Syntactical Analysis method comparison
4. Grammar with attributes
   - 4.1. Definitions
   - 4.2. YACC
5. Semantic Analysis
   - 5.1. Semantic Analysis functions
   - 5.2. Type systems
6. Symbol tables
   - 6.1. Trees
   - 6.2. Hash tables
7. Memory management
   - 7.1. Code
   - 7.2. Static
   - 7.3. Stack
   - 7.4. Heap
     - 7.4.1. Simple structures
     - 7.4.2. Marked blocks
     - 7.4.3. Buddy systems
     - 7.4.4. Garbage collection
   - 7.5. Dangling references
   - 7.6. Padding
8. Code generation
   - 8.1. Intermediate code
   - 8.2. Basic blocks
   - 8.3. Machine code
     - 8.3.1. Local strategies
     - 8.3.2. Global strategies
9. Code optimization
   - 9.1. Local optimization
   - 9.2. Global optimization
   - 9.3. Machine dependent optimization

Fig. 1. Language Processors course structure.

or 2) the variance of the estimated knowledge distribution is lower than a certain value.

Once the test has finished, the score obtained by the student is shown. For each topic involved in the test, the student's knowledge level and a histogram with his or her estimated probability distribution are provided. Finally, a diagram with the percentages of items posed from each topic and the correction of all the items of the test are supplied.

IV. The Subject of Language Processors

One of the subjects stored in SIETTE is *Language Processors*. This annual subject (divided into two semesters of approximately 14 weeks per semester) is imparted during the seventh and the eighth semesters in the Computer Science Engineering School of the University of Málaga, Málaga, Spain. The subject's teachers have been using SIETTE as a complementary academic tool since the second half of 2002, that is, during three courses. At the beginning of the course, students are provided with a username/password pair in SIETTE. All these pairs are automatically generated by SIETTE and allow teachers to identify the students. The curriculum of this subject is depicted in Fig. 1.

The primary goal of this subject is that students learn the most important issues about compilers. As a consequence, by the end of the course, students must have learned techniques about lexical, syntactical, and semantic analysis. Likewise, they must be able to construct a compiler for an imperative programming language specified by the teachers at the beginning of the second semester.

During each course, students were administered three test exams. These tests are only accessible by the subject's student group. Teachers make these tests available during the exams and restrict their access so that this access is available only by means of the PCs located in the teaching laboratories of the school.

Fig. 2 shows the aspect of an item (and its correction) of one of these tests. The window on the left shows a multiple-response item with five options in the way it is posed to a student in SIETTE. This item is composed by a stem and a set of five options. In this type of item, the student has to select all the options he or she considers correct. In addition, the student can request a hint because this item has hints available. In an item, hints are available when, below the stem, a hint button is shown. Hints are only shown on student demand. The result of pushing the hint button is that a little window (such as the one in Fig. 2) opens showing the content of the hint. In the figure example, the hint is used to clarify the meaning of the term "recognize" included in the stem. Once the student submits the answers, they are corrected. In Fig. 2, the window on the right shows how the item correction is presented to the student in SIETTE. The appearance of this Web page is very similar to the previous one. The symbols used by SIETTE at this stage are a cross to indicate the wrong answers selected by the student; a gray check mark to indicate the correct answer that has not been selected; and a black check mark to point out that the answer selected by the student is correct. In accordance with the figure example, the student selected two wrong answers (the third and the fifth ones) and did not select the only correct answer (i.e., the first one).

Psychological evidence indicates that immediate feedback after an error is the most effective pedagogical action [13]. In SIETTE, teachers can add feedback for each item answer. When added, the feedback is shown under the correction of the item (such as in Fig. 2). Feedback consists of pieces of knowledge that help students eliminate misconceptions or learn concepts unknown until that moment. There are two types of feedback [14]: 1) negative feedback, used to correct a wrong answer (e.g., a justification about why the answer is wrong) and 2) positive feedback, given to reinforce a correct answer. In the right-hand side of Fig. 2, the student has made two mistakes, and as a result, he or she has received negative feedback, shown below the item correction.

A. Evaluation

During the first semester of the 2004–2005 course, two exams of the first part of the curriculum were carried out. The first one

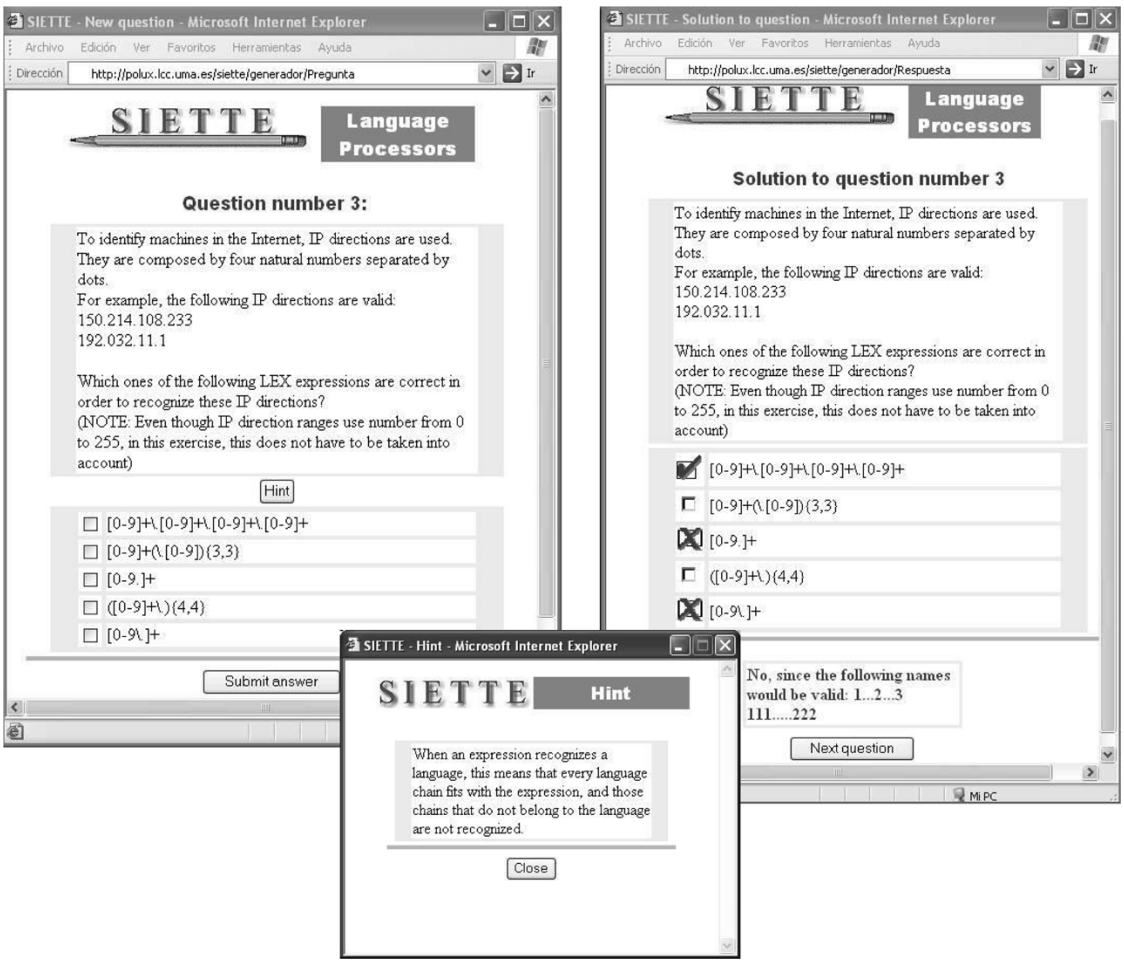

Fig. 2. Posing of an item, a hint, the item correction, and its corresponding feedback.

was a test with hints and feedback about the topic *Lexical Analysis*. This test was administered before the Christmas holidays and was taken by those students who wanted to take the test voluntarily. At the end of the semester, a second exam was administered. One of the parts of this exam was an obligatory test about the same topic but this time without hints and feedbacks and with different items than the former test. After a comparative analysis of the performances of those students that took both tests (i.e., 57 individuals), 60% of the students had improved their knowledge level. Within this percentage, 10% of the students improved their results by 50%; 25% demonstrated an improvement between 50% and 25%; 30% an improvement between 25% and 10%; and the remainder (60%) an improvement of less than 10%. Certainly these results could be biased by at least the three following factors.

1) A difference of two months existed between the administration of both tests. Thus, factors such as forgetfulness or the possibility that some students had made a deeper study of the subject must be taken into account.
2) The tests contained different item collections. An analysis of the global test difficulty revealed that the second test was more difficult.
3) Students knew *a priori* that, in contrast to the second test, the results of the first test were not determinant for the final qualification, that is, these results only would affect the final qualification if they were better than the results obtained in the second test. It is possible that, in the second test, students might feel stressed, which might affect their performances.

V. A Feasibility Study of Web-Based Adaptive Testing

SIETTE is a Web-based CAT generation system. Consequently, its tests include items whose ICCs have been previously calibrated through test session data administered nonadaptively. For this reason, in this system, the support required to administer nonadaptive tests has been included, that is, fixed-length tests assessed with conventional criteria such as the percentage of items successfully answered. Accordingly, when a subject is created, before administering CATs of its curriculum, a nonadaptive test must be given. Using the results of this test, calibration is completed, and CATs can be administered.

Although P&P testing is recognized as a good media to collect data for calibration purposes, can the same be said about Web-based testing? To answer this question, in 2002, a pilot experiment was accomplished. The goal of this experiment was to compare calibration results through a P&P test with calibration results through SIETTE. To this end, a subject of English grammar fundamentals was constructed in SIETTE. This subject could be considered as part of a K–12 curriculum for students with middle-school level. Within this subject, a test was

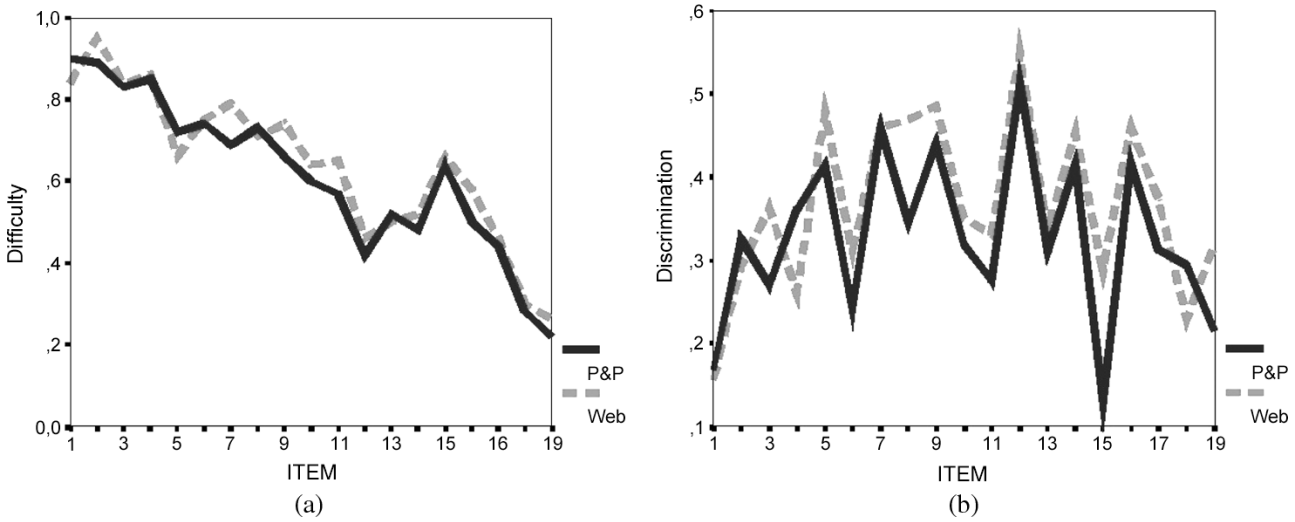

Fig. 3. Calibration result comparisons of (a) the difficulty and (b) the discrimination factor of a test item.

created. This test assessed intermediate English grammar skills and was simultaneously administered nonadaptively through SIETTE and using a P&P media. The test contained just 20 multiple-choice items with three possible answers per item. For each item, students had to complete a sentence written in the stem. The answers were phrases or just words. Students had to select the answer that best fit (grammatically) the sentence.

In SIETTE, students were recruited from different sources (via e-mail, propaganda brochures, etc.) through the Internet and three Spanish universities (the University of Málaga, the Polytechnic University of Valencia, and the Autonomous University of Madrid). Data were collected for about two months. Students taking the test were mainly from Spain and other Spanish-speaking countries.

Before the test was given to each student, a Web page with an explanation of the experiment was presented. After that, he or she had to fill in a questionnaire intended to collect some personal information about him or her: gender, age, studies, self-estimation about English level (the possible values were: practically null, low, medium, high, or practically bilingual), the origin of English language training, nationality, and mother tongue. Other information collected internally was the date and time of the connection to SIETTE, the operating system and Web browser used, and finally the exposure time spent on each item. Once this questionnaire was submitted, the instruction to take the test was shown. Each student had a time limit of 20 minutes to complete the test. When the first item was shown, the time countdown began. Items were always posed in the same order. After the test finished, the final score (percentage of correct answers) was provided.

A total number of 2316 cases were collected, but after thorough data analysis, only 1123 were considered valid for this study. The data discarded contained fully blank tests, incomplete tests with too many omissions and very short time, possible resubmissions of tests (this was detected by controlling the IP direction from which tests were taken), etc.

These data were statistically compared, and the study revealed that the P&P test and SIETTE samples differ in all the variables studied, i.e., nationality, size, sex, age, English level, and education. Items were calibrated with these data. Results showed that the difficulty of items was found to be very similar in both environments [Fig. 3(a)], although the discrimination factor varies in samples [Fig. 3(b)]. The correspondence between the parameters of the 3PL can be considered acceptable. Globally, the results obtained in both samples were very similar in terms of test performances (Fig. 4). This correspondence was more evident when the whole test was considered. Therefore, under certain conditions (i.e., in controlled environments), the calibrations made with data collected through a Web-based system are equivalent to calibrations made with data collected using a P&P media. Consequently, the Internet can be used for item calibration purposes, and accordingly, a system like SIETTE can be considered as a reliable medium for CAT construction.

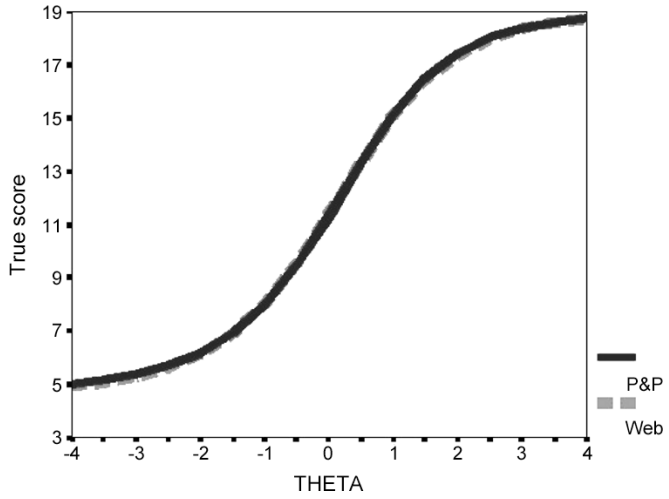

Fig. 4. Global test performance comparison.

VI. Related Work

Most Web-based testing systems are nonadaptive commercial tools. Despite being nonadaptive tools, some of them, such as Intralearn (Intralearn Software Corporation, Northboro,

MA), WebCT (WebCT, Inc., Lynnfield, MA), TopClass (WBT Systems, Dublin, Ireland), I-assess (EQL International, Ltd., Livingston, West Lothian, U.K.), and C-Quest (Assessment Systems Corporation, Helsinki, Finland), allow teachers to include only one piece of feedback per item. On the other hand, the few Web-based tools that implement CAT, such as MicroCAT (Assessment Systems Corporation, St. Paul, MN) and TerraNova CAT (CTB/McGraw-Hill, Monterey, CA), do not provide feedback and are not able to assess topics structured in a curriculum.

Other instructional Web-based systems have certain adaptive assessment features. For instance, the lisp tutor ELM-ART [15] owns a testing component that uses information about student performance in previous tests to select the next question to ask. In this system, teachers estimate item difficulty heuristically. Medtec [16] is an intelligent Web-based tutor for basic anatomy, where (nonadaptive) tests are generated automatically based on the student model. In the hypermedia learning systems presented in [17] and [18], fuzzy approaches are used to diagnose a student's answers and to create and update the student model. In both systems, the student model is used to select the problem to be presented to the student.

In general, the selection of problems in all these systems is based on the same idea: to select problems according to their difficulty and the student's performance (the better the performance, the greater the difficulty). Even such a simple strategy can result in tests that are challenging for students; but, of course, the use of IRT further exploits this possibility, achieving much better results.

VII. Conclusion and Future Work

CATs represent a revolution in conventional understanding about tests since teachers have the possibility of making individualized assessments. SIETTE is a versatile Web-based CAT generation system that provides well-founded assessment as a result of the underlying IRT. Its availability through the Internet facilitates access to the system to a large number of students. Moreover, no additional software needs to be installed other than a connection to the Internet and a Web browser tool. Multimedia capabilities can be used in Web-based systems; items with multiple formats can be easily included in tests, making them more interesting and more entertaining.

Tests represent one of the most relevant features of current learning systems [19]. In SIETTE, they can be used not only for assessment, but also for learning from the self-assessment tests. By using hints and feedbacks, students can participate in an active learning process and can contribute to improving students' knowledge and detecting some possible misconceptions. The effect of this feature has been evaluated in a subject about Language Processors imparted using SIETTE at a Spanish computer science school. Currently, in SIETTE, multiple hints per item are allowed in such a way that once a student pushes the item hint button, one of these hints is selected randomly. In the future, these hints will be adaptive. The hint shown will be the most adequate in terms of the estimated student knowledge level [14].

Before using SIETTE as a CAT generation system, its adequacy as a CAT generation platform was studied. For this purpose, an experiment with a subject of English grammar fundamentals was conducted. A nonadaptive test of this subject was administered to two different samples at the same time through SIETTE and using a P&P media. Data collected from the SIETTE sample were filtered, and invalid data were discarded. Later, item calibration was accomplished independently, on the one hand, with the results collected by P&P media, and on the other hand, with the filtered results of SIETTE. Results showed that there is no significant difference between administering a test for calibration purposes using P&P techniques and using SIETTE (in controlled environments).

At this time, the student model repository contains information about more than 15 000 test sessions, and the knowledge base of SIETTE contains 84 subjects, 1852 topics, 3820 items, and 220 tests. Most of these subjects include courses from the computer science engineering school, the telecommunication engineering school, the philosophy faculty, and a postgrade Master's of Computer Science applied to mobile communications; all of them were from the University of Málaga. For this reason, access to most of these subjects is restricted to their corresponding students. These students use SIETTE for self-assessment, although it has also been proven successful for academic assessment purposes in subjects like Language Processors. Finally, a *demo* subject exists, which contains a collection of tests designed to show the capabilities of SIETTE and its types of items. All these tests are freely available.

References

[1] H. Wainer, *Computerized Adaptive Testing: A Primer*. Hillsdale, NJ: Lawrence Erlbaum, 1990.

[2] J. R. Anderson, F. G. Conrad, and A. T. Corbett, "Skill acquisition and the lisp tutor," *Cogn. Sci.*, vol. 13, pp. 467–505, 1989.

[3] K.-E. Chang, Y.-T. Sung, K.-Y. Wang, and C.-Y. Dai, "Web_soc: A socratic-dialectic-based collaborative tutoring system on the World Wide Web," *IEEE Trans. Educ.*, vol. 46, no. 1, pp. 69–78, Feb. 2003.

[4] W. J. van der Linden and C. A. W. Glas, *Computerized Adaptive Testing: Theory and Practice*. Norwell, MA: Kluwer, 2000.

[5] R. Conejo, E. Millán, J. L. Pérez-de-la-Cruz, and M. Trella, "An empirical approach to on-line learning in SIETTE," in *Lecture Notes in Computer Science 1839. Proc. 3rd Int. Conf. Intelligent Tutoring Systems (ITS 2000)*, G. Gauthier, C. Frasson, and K. VanLehn, Eds., 2000, LNCS no. 2347, pp. 405–408.

[6] S. E. Embretson and S. P. Reise, *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum, 2000.

[7] R. K. Hambleton, J. Swaminathan, and H. J. Rogers, *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage, 1991.

[8] A. Birnbaum, "Some latent trait models and their use in inferring an examinee's mental ability," in *Statistical Theories of Mental Test Scores*, F. M. Lord and M. R. Novick, Eds. Reading, MA: Addison-Wesley, 1968.

[9] E. Guzmán and R. Conejo, "A brief introduction to the new architecture of SIETTE," in *Lecture Notes in Computer Science. Proc. 3rd Adaptive Hypermedia Adaptive Web-based Systems (AH 2004)*, Berlin, Germany, 2004.

[10] ——, "A library of templates for exercise construction in an adaptive assessment system," *Technology, Instruction, Cognition and Learning*, vol. 2, no. 1–2, pp. 21–43, 2004.

[11] ——, "Simultaneous evaluation of multiple topics in SIETTE," in *Lecture Notes in Computer Science 2363. Proc. 6th Int. Conf. Intelligent Tutoring Systems (ITS 2002)*, S. A. Cerri, G. Gouardčres, and F. Paraguacu, Eds., Berlin, Germany, 2002, pp. 739–748.

[12] R. J. Owen, "A Bayesian sequential procedure for quantal response in the context of adaptive mental testing," *J. Amer. Statistical Assn.*, vol. 70, no. 350, pp. 351–371, 1975.

[13] A. Mitrovic, "SINT—a symbolic integration tutor," in *Proc. 6th Int. Conf. Intelligent Tutoring Systems (ITS 2002), Lecture Notes in Computer Science*, S. Cerri, G. Gouardères, and F. Paraguacu, Eds., New York, 2002, LNCS no. 2363, pp. 587–595.

[14] R. Conejo, E. Guzmán, and J. L. Pérez-de-la-Cruz, "Towards a computational theory of learning in an adaptive testing environment," in *Artificial Intelligence Education: Shaping Future of Learning Through Intelligent Technologies (AIED 2003)*, U. Hoppe, F. Verdejo, and J. Kay, Eds., 2003, pp. 398–400.

[15] G. Weber and M. Specht, "User modeling and adaptive navigation support in WWW-based tutoring systems," in *Proc. 6th Int. Conf. User Modeling (UM 1997)*, A. Jameson, C. Paris, and C. Tasso, Eds., 1997, pp. 289–300.

[16] C. Eliot, D. Neiman, and M. LaMar, "Medtec: A Web-based intelligent tutor for basic anatomy," in *Proc. 2nd World Conf. WWW, Internet Intranet (WebNet 1997)*, 1997, pp. 161–165.

[17] S. H. Lee and C. J. Wang, "Intelligent hypermedia learning system on the distributed environment," in *Proc. World Conf. Educational Multimedia/Hypermedia Educational Telecommunications (ED–MEDIA/ED–TELECOM 1997)*, 1997, pp. 625–630.

[18] G. Hwang, "A test-sheet-generation algorithm for multiple assessment requirement," *IEEE Trans. Educ.*, vol. 46, no. 3, pp. 329–337, Aug. 2003.

[19] P. Brusilovsky and P. Miller, "Web-based testing for distance education," in *Proc. World Conf. WWW Internet (WebNet 1999)*, 1999, pp. 149–154.

Eduardo Guzmán received the Ph.D. degree in computer science from the University of Málaga, Málaga, Spain, in 2005.

He is an Associate Teacher at the Department of Languages and Computer Science with the University of Málaga, where he has worked since 1998. His research interests focus on adaptive testing, student knowledge diagnosis, intelligent tutoring systems, and artificial intelligence applied to civil engineering. He has been involved in the development of the SIETTE assessment system since 2000.

Ricardo Conejo received the Ph.D. degree in *ingeniero de canales caminos y puertos* from the Polytechnic University of Madrid, Madrid, Spain, in 1995.

He is an Associate Professor in the Languages and Computer Science Department of the University of Málaga, Málaga, Spain, where he has worked since 1986. In addition, he is the Director of the Research and Applications in Artificial Intelligence Group at the same university. He has 15 years of experience in the development of intelligent tutoring systems. He has published extensively on this field. His research interests currently focus on adaptive testing, student knowledge diagnosis, intelligent tutoring systems, and multiagent systems. He has also worked on fuzzy logic, model-based diagnosis, and artificial intelligence applied to civil engineering.

Dr. Conejo is a regular member of program committees of international conferences, such as Intelligent Tutoring Systems (ITS), Adaptive Hypermedia and Adaptive Web-Based Systems (AH), and Artificial Intelligence in Education (AIED).