Int J Artif Intell Educ (2016) 26:270–292
DOI 10.1007/s40593-015-0078-4

ARTICLE

The SIETTE Automatic Assessment Environment

Ricardo Conejo[1] · Eduardo Guzmán[1] ·
Monica Trella[1]

Published online: 10 November 2015

Abstract This article describes the evolution and current state of the domain-independent Siette assessment environment. Siette supports different assessment methods—including classical test theory, item response theory, and computer adaptive testing—and integrates them with multidimensional student models used by intelligent educational systems. Teachers can use an authoring tool to create large item pools of different types of questions, including multiple choice, open answer, generative questions, and complex tasks. Siette can be used for formative and summative assessment and incorporates different learning elements, including scaffolding features, such as hints, feedback, and misconceptions. It includes numerous other features covering different educational needs and techniques, such as spaced repetition, collaborative testing, or pervasive learning. Siette is designed as a web-based assessment component that can be semantically integrated with intelligent systems or with large LMSs, such as Moodle. This article reviews the evolution of the Siette system, presents information on its use, and analyses this information from a broader and critical perspective on the use of intelligent systems in education.

Keywords Automatic assessment · Item Response Theory · Computerized adaptive testing

Introduction

The role of assessment is essential to measuring the learners' achievements during the teaching and learning processes. It is also clear that the learners' current state of

✉ Ricardo Conejo
conejo@lcc.uma.es

Eduardo Guzmán
guzman@lcc.uma.es

Monica Trella
trella@lcc.uma.es

[1] Universidad de Málaga, 29071 Málaga, Spain

knowledge must be known in order to adapt instruction to the learners' needs. Assessment also plays an important role in meta-cognition and motivation, for example. Classic Intelligent Tutoring Systems (ITSs) (Polson and Richardson 2013) have been developed that attempt to emulate the teacher's behaviour with an individual learner and to tailor instruction to each learner's circumstances. The core of the ITS is a computable representation of the learner called the learner model (LM) (Greer and McCalla 1994). The LM can include different features, but the domain structure, the learning goals, and the current state of knowledge is almost always represented.

Traditionally, the learner knowledge models used by ITSs are complex representations that include nodes corresponding to each concept or set of concepts in the domain. These nodes usually have a qualitative or quantitative value attached that indicates the knowledge level for that particular concept. Additionally, some ITSs can include nodes that represent misconceptions and arrows that link concepts. These links represent relationships used to make inferences about the knowledge level for some concepts according to the estimated knowledge level for other concepts. A classic approach in the ITS field is to heuristically define the concepts, relationships, and propagation rules.

Assessment in learning, understood as a research field whose goal is to accurately measure educational achievement, has also evolved separately as part of the discipline of psychometrics. The 20th century saw the emergence of relevant contributions to the field, such as the *Classical Test Theory* (CTT) (Lord et al. 1968), the *Item Response Theory* (IRT) (Embretson and Reise 2000), and the *Computer Adaptive Testing* (CAT) (Wainer et al. 2000). All these theories have important strengths: they are domain independent, data-driven, mathematically well-founded and, due to the two previous features, the parameters used to tune the model can be estimated by statistical data analysis.

By the end of 20th century, the field of ITS produced its own domain-independent models or techniques, such as *Knowledge Tracing* (KT) (Corbett and Anderson 1994), *Constraint Based Modelling* (CBM) (Mitrovic 2012), and some others. However, a question remained unanswered: Why did ITS researchers not use psychometric models? Several answers are possible. The most straightforward answer is that ITS requires data that is more fine-grained than the information usually provided by psychometric models, i.e. unidimensional measures, and that psychometric models typically require complete data sets and lots of response data to be applied. Another possible reason is that psychometric models have been classically linked to Multiple-Choice Questions (MCQs) and most tasks involved in ITSs require *High-Order Cognitive Skills* (HOCS) (e.g. problem solving), which are difficult to assess using MCQs. Finally, the most commonly used psychometric models assume that no learning can happen during the assessment process.

The question stated in the foregoing drove the initial research that motivated the development of the Siette system (http://www.siette.org). The primary aim was to create a practical flexible system that uses state-of-the-art methods from research on psychometrics and ITSs to facilitate the creation and administration of formative and summative assessments. It has also been a test bed for different features related to automatic assessment. Siette is an ongoing long-term project that has both improved and deepened our understanding of the role of assessment in education. All these features have been added to the Siette architecture and work in a fully integrated manner. The article is organised as follows: the next section briefly describes the origins and evolution of Siette; we then discuss how different problems have been addressed during its development and how their solutions have been included in the system as new features; next, we provide statistical data on the use of Siette; finally, some conclusions are presented.

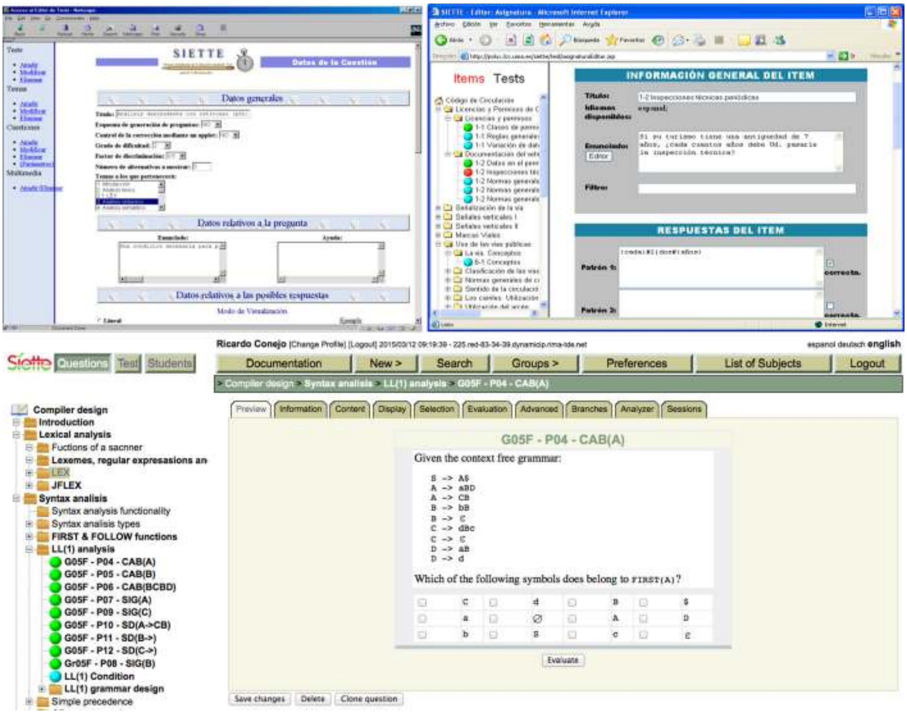
A Brief History of Siette

Siette has evolved during 16 years of practical use. The current system varies considerably from the initial version. Some features have been described in separate studies and integrated within the Siette architecture, whereas other features have been added but not yet described because the appropriate experiments have not yet been conducted.

Siette was initially developed as the focus of a master thesis (Rios 1998). The initial aim was simply to use the advantages of web interfaces (whose popularity began to significantly increase at that time) to replace classic paper and pencil tests, which were widely used in many undergraduate courses. We explored the field and encountered IRT and CAT. IRT is based on the hypothesis that student knowledge can be measured as a single real number. Conditional probabilities functions are defined to explain the response to the questions (called items). The higher the knowledge level, the higher the probability of solving the questions. Statistical procedures can be used to infer the students' knowledge level based on their actual responses. On the other hand, CAT tries to modify the selection of questions to maximize the information obtained. It can be proved that this condition is equivalent to the selection of the question whose difficulty is closer to the currently estimated student knowledge.

We realize that the Web was a perfect platform to deliver tests within the framework of these two theories. By that time, we were working in a European project called TREE (Trella et al. 2000), which included the development of an expert system and a web-based ITS for the identification and classification of European trees. We planned to use Siette as a component of the TREE project to automatically generate questions from the project database and feed the system LM with these data (Rios et al. 1998, 1999; Conejo et al. 2004). The initial idea evolved from it being a standalone system to being a component of a larger system. However, Siette was always designed to be a reusable and domain-independent module.

One of our challenges was to develop a web-based system that could not only be used for research purposes, but could also be used in real environments, i.e. with real learners and teachers. However, this challenge involved certain demands, among which were the following three aspects that had always been considered during the development of Siette. Firstly, *the system had to be robust and efficient*. This demand involved considerable effort in software engineering. Three main versions of the system have been developed. The first, which was described by Conejo et al. (2004), was implemented in PHP and pure HTML, with CGI technology written in the C language. The second version (Guzmán and Conejo 2004b, c), which was developed from scratch, and the third version (Guzmán et al. 2007b) were developed in Java with Servlets, JSP, Javascript, Ajax, and some other web technologies. Secondly, *the system had to provide a user-friendly authoring tool* (Guzmán et al. 2005). This was the main difference between the second and third versions. Figure 1 shows the evolution of the Siette authoring tool.


SIETTE
Datos de la Cuestion
Datos generales
Datos relativos a la pregunta
Enunciado:
Ayuda:
Datos relativos a las posibles respuestas
Modo de Visualización
SIETTE - Editor: Asignatura - Microsoft Internet Explorer
Items Tests
Código de Circulación
INFORMACIÓN GENERAL DEL ITEM
RESPUESTAS DEL ITEM
Siette Questions Test Students
Documentation New > Search Groups > Preferences List of Subjects Logout
espanol deutsch english
Preview Information Content Display Selection Evaluation Advanced Branches Analyzer Sessions
Evaluate
Save changes Delete Clone question

Titulo: 1-2 Inspecciones técnicas periódicas
Idiomas disponibles: español
Enunciado: Si su turismo tiene una antigüedad de 7 años, ¿cada cuantos años debe Ud. pasarle la inspección técnica?
Patrón 1: (cada|#2|dos#|años) ☑ correcta.
Patrón 2: □ correcta.

Ricardo Conejo [Change Profile] [Logout] 2015/03/12 09:19:39 - 225.red-83-34-39.dynamicip.rima-tde.net

> Compiler design > Syntax analisis > LL(1) analysis > G05F - P04 - CAB(A)

- Compiler design
  - Introduction
  - Lexical analysis
    - Fuctions of a sacnner
    - Lexemes, regular expresasions an
    - LEX
    - JFLEX
  - Syntax analisis
    - Syntax analysis functionality
    - Syntax analisis types
    - FIRST & FOLLOW functions
    - LL(1) analysis
      - G05F - P04 - CAB(A)
      - G05F - P05 - CAB(B)
      - G05F - P06 - CAB(BCBD)
      - G05F - P07 - SIG(A)
      - G05F - P09 - SIG(C)
      - G05F - P10 - SD(A->CB)
      - G05F - P11 - SD(B->)
      - G05F - P12 - SD(C->)
      - Gr05F - P08 - SIG(B)
      - LL(1) Condition
      - LL(1) grammar design
  - Simple precedence

G05F - P04 - CAB(A)

Given the context free grammar:

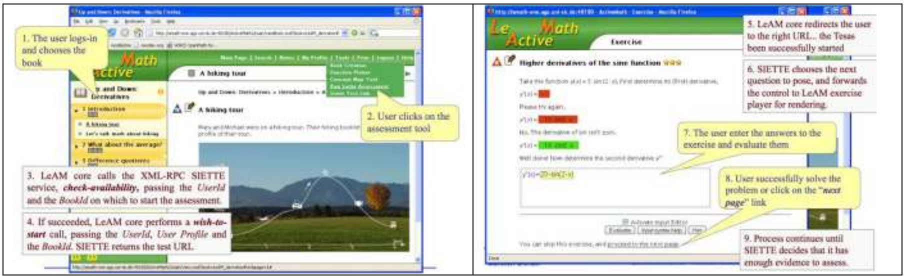
```
S -> A$
A -> aBD
A -> CB
B -> bB
B -> ε
C -> dBc
C -> ε
D -> aB
D -> d
```

Which of the following symbols does belong to FIRST(A)?

□	c	□	d	□	B	□	$
□	a	□	∅	□	A	□	D
□	b	□	S	□	C	□	ε

Fig. 1 Evolution of the Siette authoring tool

Finally, *Siette had to be interoperable with other systems*. To fulfil this requirement, we defined a protocol based on web services that allowed Siette to be used remotely. These web services initially enabled the connection with ActiveMath system (Melis et al. 2001) as part of our participation in the European project LeActiveMath (http://leactivemath.org). Figure 2 shows the sequence of actions during the LeActiveMath-Siette communication. Subsequently, these services were refactored to support new kinds of items (described in the next section) and an API, which was developed in PHP, in order to smoothly integrate Siette and Moodle LMS. Moodle users (teachers and students) can use Siette as another Moodle activity, while preserving the role they have in Moodle (as teachers or students). The results obtained in Siette are synchronously passed to the LMS. Siette is currently integrated with the Moodle used as a virtual



1. The user logs-in and chooses the book

2. User clicks on the assessment tool

3. LeAM core calls the XML-RPC SIETTE service, ***check-availability***, passing the *UserId* and the *BookId* on which to start the assessment.

4. If succeeded, LeAM core performs a ***wish-to-start*** call, passing the *UserId*, *User Profile* and the *BookId*. SIETTE returns the test URL.

5. LeAM core redirects the user to the right URL... the Tesas been successfully started

6. SIETTE chooses the next question to pose, and forwards the control to LeAM exercise player for rendering.

7. The user enter the answers to the exercise and evaluate them

8. User successfully solve the problem or click on the "***next page***" link

9. Process continues until SIETTE decides that it has enough evidence to assess.

Fig. 2 Integration with LeActiveMath environment

campus in all the schools and faculties of Malaga University (2500 staff; 35,000 students per year) Fig. 3 shows a Moodle form to define a Siette assessment definition, and a test is ready to be taken.

Problem Addressed and Current Features of Siette

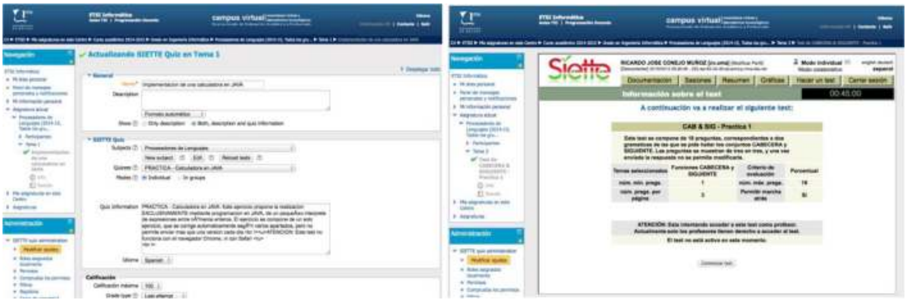
Siette has changed from being a test-based system to a current environment for automatic assessment. In this section, we summarize the most relevant features of Siette to provide readers with an overview of the current functionalities provided by the system. Over the years, Siette has faced several problems that have been turned into challenges. Their solution has led to the system becoming enriched by the addition of different features, as described below. The next section addresses the issue of the adoption of these features by users.

The Domain Model

Siette is a domain-independent system for automatic assessment. A subject or domain in Siette is structured hierarchically ("part-of" relationships), with topics and subtopics. Each subject contains an item pool in which items can be attached to any node in the hierarchy, indicating that knowledge on that concept is required to solve that item (unidimensional model). The system also supports linking items to two or more sibling nodes of the hierarchy, under the constraint that those nodes have to be siblings (multidimensional model); however, multidimensional items are rarely used.

Additionally Siette allows the "prerequisite-of" relationship between nodes of the model. Although it is not currently used for assessment purposes, it is used improve the graphical model presentation. In addition to the domain model, a hierarchically structured misconception model can be defined. Incorrect responses to items can also be associated with any of their nodes (Guzmán and Conejo 2015).

The Siette domain model is designed in this way for two reasons: (1) To allow integration with ITS hierarchical domain models. Siette is not an ITS, but an assessment system that can be integrated into an ITS, so it uses a domain model that can be overlaid with the ITS domain model; (2) To enhance adaptivity based on content selection. In addition to classic CAT behaviour, the Siette adaptive question-selection algorithm takes into account that multiple domain model nodes



Actualizando SIETTE Quiz en Tema 1

General

Description

Show: ○ Only description ● Both, description and quiz information

SIETTE Quiz

Subjects: Procesadores de Lenguajes

New subject | Edit | Reload tests

Quizzes: PRACTICA - Calculadora en JAVA

Modes: ● Individual ○ In groups

Idioma: Spanish

Calificación

RICARDO JOSE CONEJO MUÑOZ

Modo individual

Información sobre el test 00:45:00

A continuación va a realizar el siguiente test:

CAB & SIG - Practica 1

ATENCIÓN: Esta intentando acceder a este test como profesor. Actualmente solo los profesores tienen derecho a acceder al test.

El test no está activo en este momento.

Fig. 3 Integration with Moodle at the Malaga University Virtual Campus

can be evaluated in a single assessment session and selects the next question to pose according to the nodes that require higher precision (Guzmán et al. 2007a).

The Learner Model and the Problem of Multidimensionality

Multidimensionality was the main problem when we tried to use a CAT as an assessment component of an ITS. ITSs are commonly based on a fine-grained LM (McCalla and Greer 1994), which consists of different knowledge estimates for each domain topic. However, CTT and IRT generally focus on a single latent trait and produce a single numerical value as an estimate of aggregate knowledge. Multidimensional IRT models were quite complex and more than two or three dimensions are rarely used. Classic multidimensional models are based on parametric families of functions for the item characteristic curves. In the early days of Siette we adopted a non-parametric multidimensional model and explored kernel-smoothing techniques (Ramsay 1991). However, for practical reasons, we adopted a multi-unidimensional approach and developed a mechanism that assumes that multiple independent traits are assessed simultaneously. The implications associated with this assumption and the problem of aggregating concepts scores in a hierarchical LM are described in (Guzmán and Conejo 2002; Guzmán et al. 2007a). Currently, Siette does not support true multidimensional items, but this line might be taken up again in the future.

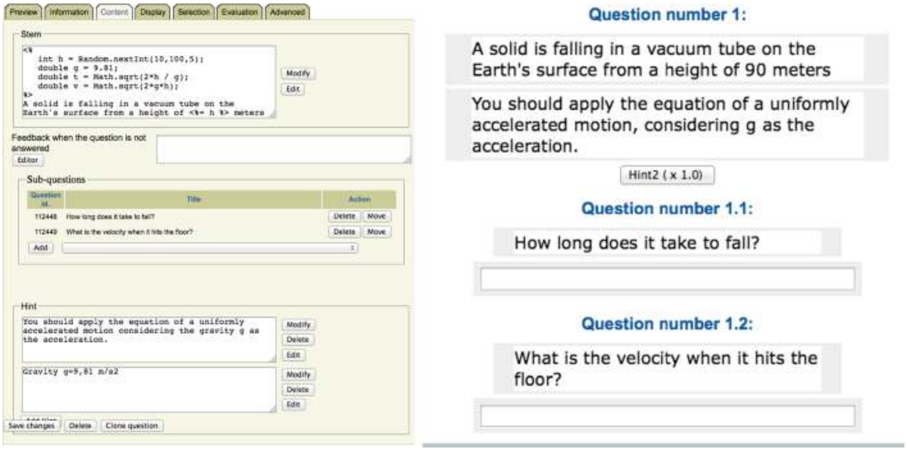
We have also explored from a theoretical point of view the relationship between qualitative and qualitative LMs that are commonly used in ITS and IRT, respectively, and how unidimensional and multidimensional models can be related based on the structure imposed by the prerequisite relationship (Pérez-de-la-Cruz et al. 2005).

Item Models and Types

One of the keys to the practical success of Siette is the possibility of using a combination of different item types in the same test. In the initial version items were MCQs, but in the current version an item can be considered to be any component able to provide information on student knowledge. Every item usually contains a stem, a set of answers, a set of hints, and feedback. Both hints and feedback are optional and can be shown to the student depending on the test configuration parameters. Siette supports *basic and complex item types*. Moreover, the system is able to assess problem-solving skills and is capable of automatic item generation, as described below. See Figs. 4, 5, 6, and 7.

Basic Item Types There are three basic item types in Siette: *multiple-choice questions with a single answer*, *multiple-choice questions with multiple answers*, and *open multiple short-answer questions*.

Open multiple short-answer items can be configured to support a reduced set of texts as item answers by providing patterns (i.e. regular expressions) for evaluating item correction. The answer texts are matched against a set of given patterns using pattern evaluation components. The easiest component only matches the correspondence between the answer and the pattern, ignoring blanks, upper and lower case letters, punctuation signs, etc. Furthermore, Siette includes a pattern evaluation component that matches regular expressions, number ranges, units, and magnitudes conversion, etc. For



Stem

```
<%
   int h = Random.nextInt(10,100,5);
   double g = 9.81;
   double t = Math.sqrt(2*h / g);
   double v = Math.sqrt(2*g*h);
%>
A solid is falling in a vacuum tube on the
Earth's surface from a height of <%= h %> meters
```

Feedback when the question is not answered

Sub-questions

Question Id.	Title	Action
112448	How long does it take to fall?	Delete Move
112449	What is the velocity when it hits the floor?	Delete Move

Hint

```
You should apply the equation of a uniformly
accelerated motion considering the gravity g as
the acceleration.
```

```
Gravity g=9,81 m/s2
```

Question number 1:

A solid is falling in a vacuum tube on the Earth's surface from a height of 90 meters

You should apply the equation of a uniformly accelerated motion, considering g as the acceleration.

Question number 1.1:

How long does it take to fall?

Question number 1.2:

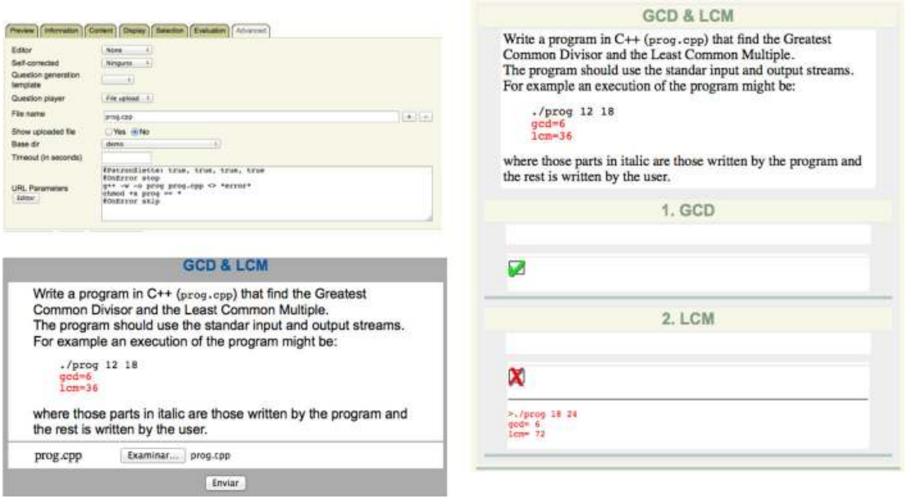
What is the velocity when it hits the floor?

Fig. 4 Generative composite items with hints

example, the system could recognize a student answer "v=70 km/h" with a given pattern "*#20 m/s#5 %". Siette has other special-purpose evaluation components, such as the ability to recognize the equivalence of a single variable function for math calculus. The Siette authoring tool provides teachers with a "re-assessment" feature that allows them to change the recognition patterns and to re-evaluate the learners' previous test session that contains those items.

Complex Item Types From the point of view of assessment, any other type of item fits within one of the three basic models described above. The following item types are in this category:

- *Siettlets*. These items were developed for tasks requiring high user interaction and were implemented using Java applets and embedded Javascript code (Arroyo et al.



Editor: None
Self-corrected: Ninguna
Question generation template:
Question player: File upload
File name: prog.cpp
Show uploaded file: ○ Yes ● No
Base dir: demo
Timeout (in seconds):
URL Parameters:

```
#PatronSiette: true, true, true, true
#OnError stop
g++ -w -o prog prog.cpp <> *error*
chmod +x prog == *
#OnError skip
```

GCD & LCM

Write a program in C++ (`prog.cpp`) that find the Greatest Common Divisor and the Least Common Multiple.
The program should use the standar input and output streams.
For example an execution of the program might be:

```
./prog 12 18
gcd=6
lcm=36
```

where those parts in italic are those written by the program and the rest is written by the user.

prog.cpp prog.cpp

GCD & LCM

Write a program in C++ (`prog.cpp`) that find the Greatest Common Divisor and the Least Common Multiple.
The program should use the standar input and output streams.
For example an execution of the program might be:

```
./prog 12 18
gcd=6
lcm=36
```

where those parts in italic are those written by the program and the rest is written by the user.

1. GCD

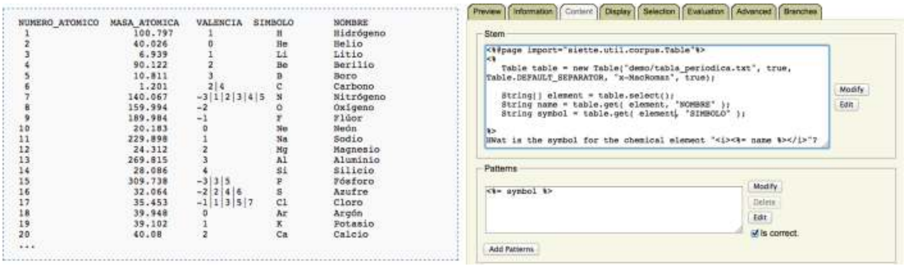
✓

2. LCM

☒

```
>./prog 18 24
gcd= 6
lcm= 72
```

Fig. 5 An external item that requires the student to upload a file



```
NUMERO_ATOMICO  MASA_ATOMICA   VALENCIA  SIMBOLO       NOMBRE
  1                100.797      1          H           Hidrógeno
  2                 40.026      0          He          Helio
  3                  6.939      1          Li          Litio
  4                 90.122      2          Be          Berilio
  5                 10.811      3          B           Boro
  6                  1.201      2|4        C           Carbono
  7                140.067     -3|1|2|3|4|5 N          Nitrógeno
  8                159.994     -2          O           Oxígeno
  9                189.984     -1          F           Flúor
 10                 20.183      0          Ne          Neón
 11                229.898      1          Na          Sodio
 12                 24.312      2          Mg          Magnesio
 13                269.815      3          Al          Aluminio
 14                 28.086      4          Si          Silicio
 15                309.738     -3|3|5      P           Fósforo
 16                 32.064     -2|2|4|6    S           Azufre
 17                 35.453     -1|1|3|5|7  Cl          Cloro
 18                 39.948      0          Ar          Argón
 19                 39.102      1          K           Potasio
 20                 40.08       2          Ca          Calcio
 ...
```

Stem

```
<%@page import="siette.util.corpus.Table"%>
<%
   Table table = new Table("demo/tabla_periodica.txt", true,
Table.DEFAULT_SEPARATOR, "x-MacRoman", true);

   String[] element = table.select();
   String name = table.get( element, "NOMBRE" );
   String symbol = table.get( element, "SIMBOLO" );

%>
WHat is the symbol for the chemical element "<i><%= name %></i>"?
```

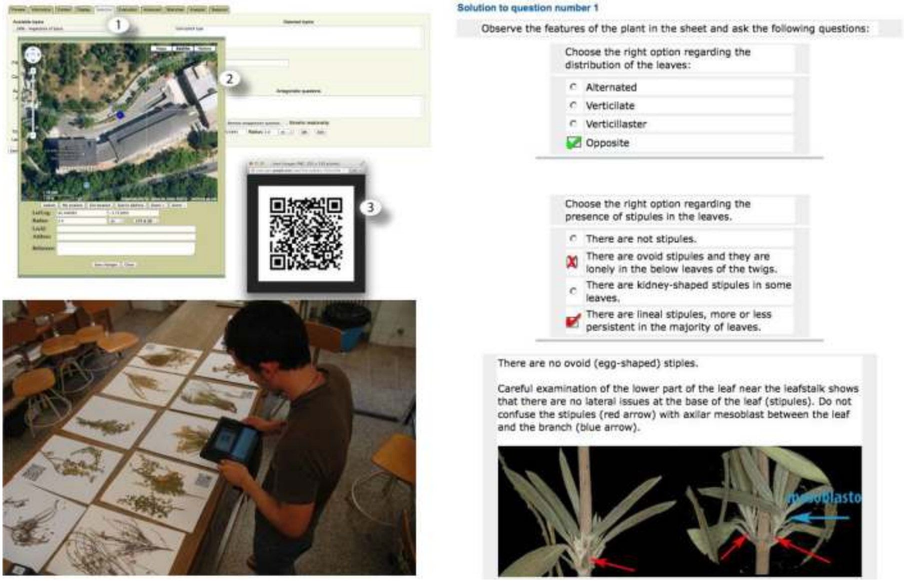
Patterns

```
<%= symbol %>
```

Fig. 6 Automatic question generation from a table

2001). A reusable library of item types was developed using this approach (Guzmán and Conejo 2004a). The mechanism developed for this purpose facilitates the integration of external complex exercises or tasks, which were the precursors of the external items (described below).

- *Composite items*. Single items can be grouped into a *composite item*. Thus, the system has to simultaneously present these items, with a common stem. Composite items play an important role in including external programs for complex task evaluation. In the literature on testing, these kinds of items are commonly known as *testlets*. Figure 4 shows an example of a physics question that is generated from a template (see the Item Generation subsection below). Questions can include hints and feedback.
- *Branching items*. Siette allows the definition of *branching items* and *branching tests*. This is an earlier technique for adaptive behaviour that involves the presentation of some items in a given order, but it is also of interest for certain applications, such as simulating question-answer dialogues or Socratic tutoring.



Solution to question number 1

Observe the features of the plant in the sheet and ask the following questions:

Choose the right option regarding the distribution of the leaves:

○ Alternated
○ Verticilate
○ Verticillaster
✓ Opposite

Choose the right option regarding the presence of stipules in the leaves.

○ There are not stipules.
☒ There are ovoid stipules and they are lonely in the below leaves of the twigs.
○ There are kidney-shaped stipules in some leaves.
✓ There are lineal stipules, more or less persistent in the majority of leaves.

There are no ovoid (egg-shaped) stiples.

Careful examination of the lower part of the leaf near the leafstalk shows that there are no lateral issues at the base of the leaf (stipules). Do not confuse the stipules (red arrow) with axilar mesoblast between the leaf and the branch (blue arrow).

Fig. 7 Assessment of botany at the laboratory using dried plants

- *Antagonistic items.* Siette permits the definition of antagonistic items, i.e., items that cannot simultaneously appear in the same test. This feature is useful when defining automatic rules for item selection in a large item pool. In the near future, the automatic detection of antagonistic items will be included in Siette according to point biserial correlation data.
- *External items.* Through this new type of item, any web-based assessment component can be integrated into Siette using a loosely coupling protocol. External items must internally correspond to one or more basic items in Siette. When one of these external items is posed to the student, Siette leaves execution control to the item. Once the student has finished interacting with the item, control is given back to Siette and the answers are provided. For instance, using this technique, Siette can call a system that evaluates the correctness of a programming exercise and returns the result to Siette in the form of a set of responses to a Siette composite item or a set of items. The next subsection describes the key role that this type of item plays in the current version of the system.

High-Order Cognitive Skills Assessment The assessment of HOCS is an important problem that we are still addressing with Siette. External items were added in order to incorporate problem-solving tasks. Furthermore, we have recently added composite items which, in combination with external items and CBM techniques, have allowed us to assess declarative knowledge (Gálvez 2009, 2012; Gálvez et al. 2009a, b, c, 2010, 2012, 2013) in problem-solving items. In parallel, and also using external and composite items, we have explored new ways of assessing procedural knowledge using IRT (Hernando 2011; Hernando et al. 2013a, b). As a consequence, the new functionality provided by these new types of items provides more flexibility in the assessment of HOCS. Figure 5 shows an example question that requires the student to write a small program. When the question is posed, the student has to upload a file containing the program that is automatically corrected according to a script provided by the teacher using the Siette authoring tool.

Item Generation Since the first version of Siette was developed, it has included the capability of generating questions using a template, for which Siette uses a simple but very effective technique: The template is written in any language script supported by the web server and the page is randomly instantiated when presenting the question to the student. Siette currently supports the addition of PHP or JSP code in the item templates. This script code can be inserted in the item stem, the responses or patterns, the hints, and the feedback. Siette provides a built-in JSP API to facilitate programming the template. This API also supports the automatic generation of questions from dictionaries and 2D tables. Figure 6 shows an example of a question generated from the chemistry periodic table. JSP programs can be defined to generate different questions based on the table values. An API is provided to make the programming task easier for non-programmers (also see Fig. 4). Extending this idea, Siette also support question generation from database tables or views. Currently we are working on question generation based on Semantic Web SPARQL queries.

Assessment Models and Test Assembly Criteria

Classical Test Theory and IRT are the theoretical assessment models underlying Siette. Siette uses an IRT-based 4PL response model including item difficulty, discrimination, guessing, and slip parameters. This model can be easily adapted to the classic 1, 2, and 3PL IRT approaches. In addition, the three basic item types are supported, respectively, by three IRT-based polytomous item models (Guzmán and Conejo 2005a; Guzmán 2005). That is, items are not always assessed dichotomously (correct or incorrect). Each item choice can be treated separately providing partial credit. This feature provides richer information and even evidence of potential misconceptions (Guzmán et al. 2010). All these options can be configured according to the teacher's specific needs.

As mentioned, Siette implements a discretization of unidimensional and multidimensional IRT models. Furthermore, we have recently added the capability of using classic continuous unidimensional IRT. The teacher can decide which assessment model to select while constructing the test specification. Many other aspects can be configured in a test specification. For instance, tests can be assembled from the item pool by either selecting items randomly, selecting them in a given order, keeping a given proportion of related topics, or using different adaptive criteria (the most informative item, the one closest to the learner's current state of knowledge, etc). Test finalization criteria can also be defined according to several options, such as fixed test length, fixed accuracy, maximum time, and the learner's own decision.

Several other parameters can be configured, such as the display style, test access options, test navigation (allowing or prohibiting forward and backward movement while solving the test items), time constraints, item exposure constraints, and the way the correct answer is presented after finishing the test.

The Hypothesis of Constant Knowledge and the "Assessment of Learning" Paradigm

Most psychometric models assume that there is no variation in knowledge levels during assessment. This hypothesis is assumed in order to allow for factoring a large joint probability density into a product of factors during estimation and for measurement error minimization techniques to be applied. This hypothesis is in clear opposition to the main goal of the ITS and other environments that are designed for learning. If Siette is used as a component of an ITS, or if it incorporates some hints and feedback during assessment, is it still valid to assume this hypothesis?

In general, this problem has been avoided by assuming that assessment only provides a snapshot of the learner's knowledge at a given time, thus upholding the hypothesis of non-learning during assessment. However, we have explored the use of assessment as a learning resource in Siette. In this case, the main goal is not only to provide an accurate score of knowledge levels on a variety of concepts or topics, but also to improve the learners' knowledge by their taking a test. For this purpose, we have used various techniques that are described below. Additionally, some indicators have been added to Siette to measure the learning gains in these cases. We are also working on more accurate models that take learning during assessment into account.

- *Adaptive item hints and feedback*. In Siette, hints are defined as pieces of information that the learners might receive before they answer an item. On the other hand, feedback consists of pieces of information that are given to the learners after they have responded. Several hints can be associated with the same item, but only one can be presented to the learner at the same time as the item. For items with more than one hint, different strategies for selecting the one that will be presented to the learner can be configured through the test configuration parameters. These strategies include randomly presenting the hint and presenting it according to current knowledge. Furthermore, feedback can be linked to each item answer. Siette can also be used as an adaptive tutoring tool as it provides hints and feedback (Conejo et al. 2005, 2006).
- *Self-assessment tests*. These turn the system into a drill and practice environment in which students can attempt the same test many times under different conditions. We have conducted several studies that suggest that these kinds of tests are of benefit to the learner (Guzmán and Conejo 2005b; Guzmán et al. 2007b).
- *Spaced-repetition test*. This is a test taken repeatedly by a user and adapts the probability of question selection according to success or failure during a previous session.
- *Collaborative testing*. These are tests that are taken simultaneously by a group of two or more learners. During these tests, learners can discuss their answers using a chat tool and, consequently, learn from their peers (Barros et al. 2007; Conejo et al. 2008, 2009a, b, 2013).

Pervasive Learning

Siette allows attaching a location tag or a QR code to each question. In this case, the question is triggered when the location is reached or the QR-code is scanned. This feature opens up the possibility of using Siette as a ubiquitous assessment tool on mobile devices (Conejo et al. 2015). Figure 7 shows an application. QR codes are assigned to questions using the authoring tool and the codes are attached to a dried plant sheet. Assessment is conducted at the botany laboratory. Questions about the plant are posed and the student receives detailed feedback after his/her response.

The Problem of Item Calibration

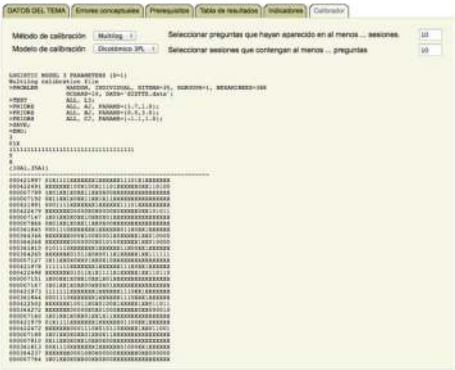
The correct implementation of Item Response Theory and the performance of CAT depend on the accurate calibration (i.e. the data-driven procedure for tuning the model parameters) of the item (question) parameters. These parameters should be computed before being used, which requires large datasets. The only alternative method is the heuristic estimation of the parameters. We analysed the sensitivity of the model under parameter mis-estimation (Conejo et al. 2000) and the accuracy of heuristically estimating several important parameters, such as item difficulty, by teachers and students (Conejo et al. 2014). Evidence collection is a challenge in psychometric models and in assessment (Mislevy and Riconscente 2006). As mentioned, in order to correctly calibrate an IRT model, no learning can occur during a test session. In uncontrolled environments, such as web-based systems, a major problem is the quality of the datasets used for calibration. We have also studied the conditions under which the data collected through the Web can be used as a reliable source for calibration (Guzmán et al. 2000, 2005).

Siette performs item calibration through web services. This approach allows decoupling between the system and the calibration tool. Currently, two calibration tools are used: Multilog and JICS. Multilog is one of the most well-known tools for item calibration and has been integrated in Siette using a wrapper, whereas JICS (Java Item Calibration System) is a calibration tool that we developed independently. We have also explored other calibration mechanisms based on kernel smoothing techniques (Guzmán 2005; Guzmán and Conejo 2005a; Guzmán et al. 2007a). Figure 8 shows a preview of data prepared to call the Multilog Web Service for a set of items. The image on the right shows the returning values ready to be updated in the Siette database.

On the other hand, many Siette tests are developed by university instructors and administered to relatively small samples of students. In these situations, automatic item calibration is not always possible due to convergence issues or problematic parameter outputs. The alternative is to estimate the parameters and manually enter them into the system. This estimation can also be biased and not fully reliable. In fact, some studies have indicated that, in some cases, students can estimate some parameters better than instructors (Conejo et al. 2014). Nevertheless, in these cases, the most frequently applied practical solution is simply not to use IRT. Percentages of correct responses or scoring procedures are implemented as alternative assessment criteria. We believe that this is the reason for the relatively infrequent use of IRT scoring procedures in Siette (see Section The Use of Siette).

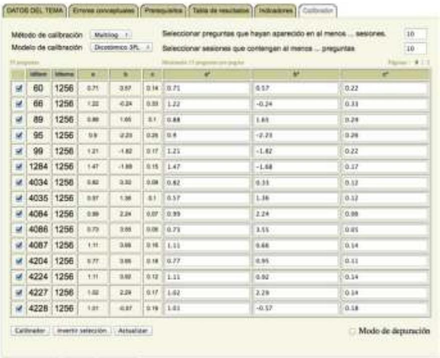
Analysis of Item and Test Results

Siette includes a built-in package for test result analysis. The learners have access to their own model and teachers have access to individual and group data. In addition to statistical descriptive data, some test consistency indicators are automatically calculated, such as Cronbach's alpha, Gutman's lambda, and tetrachoric correlation matrix. Siette also provides built-in tools for item behaviour analysis, such as point biserial correlation and item characteristic curves. Whenever possible, each value is reported



Método de calibración: Multilog
Modelo de calibración: Dicotómico 3PL
Seleccionar preguntas que hayan aparecido en al menos ... sesiones: 10
Seleccionar sesiones que contengan al menos ... preguntas: 10

```
LOGISTIC MODEL 3 PARAMETERS (0-1)
Multilog calibration file
>PROBLEM        RANDOM, INDIVIDUAL, NITEMS=15, NGROUPS=1, NEXAMINEES=388
                NCHARS=10, DATA='SIETTE.data';
>TEST           ALL, L3;
>PRIORS         ALL, AJ, PARAMS=(1.7,1.0);
>PRIORS         ALL, BJ, PARAMS=(0.0,3.0);
>PRIORS         ALL, CJ, PARAMS=(-1.1,1.0);
>SAVE;
>END;
2
01X
111111111111111111111111111111111111
Y
8
(10A1,15A1)
=====================================================
[illegible]
```



Método de calibración: Multilog
Modelo de calibración: Dicotómico 3PL
Seleccionar preguntas que hayan aparecido en al menos ... sesiones: 10
Seleccionar sesiones que contengan al menos ... preguntas: 10

	idItem	idTema	a	b	c	a'	b'	c'
☑	60	1256	0.71	0.57	0.14	0.71	0.57	0.22
☑	66	1256	1.22	-0.24	0.33	1.22	-0.24	0.33
☑	89	1256	0.88	1.65	0.1	0.88	1.65	0.29
☑	95	1256	0.9	-2.23	0.26	0.9	-2.23	0.26
☑	99	1256	1.21	-1.82	0.17	1.21	-1.82	0.22
☑	1284	1256	1.47	-1.68	0.15	1.47	-1.68	0.17
☑	4034	1256	0.82	0.33	0.08	0.82	0.33	0.12
☑	4035	1256	0.57	1.36	0.1	0.57	1.36	0.12
☑	4064	1256	0.99	2.24	0.07	0.99	2.24	0.08
☑	4066	1256	0.73	3.55	0.08	0.73	3.55	0.05
☑	4067	1256	1.11	0.66	0.16	1.11	0.66	0.14
☑	4204	1256	0.77	0.95	0.18	0.77	0.95	0.11
☑	4224	1256	1.11	0.92	0.12	1.11	0.92	0.14
☑	4227	1256	1.02	2.29	0.17	1.02	2.29	0.14
☑	4228	1256	1.01	-0.57	0.19	1.01	-0.57	0.18

□ Modo de depuración

Fig. 8 Calibration tools

with a standard 95 % confidence interval or using a colour code that indicates statistical confidence. Figure 9 shows two examples of the analysis tools. The image on the left shows the tetrachoric matrix of a set of items. The colors green and red represent positive and negative correlations, respectively, and grey represents statistically non-significant values. The right image shows the item characteristic curves for the six choices of a question, including the option to leave it blank. These features allow the teachers to identify items that have been incorrectly constructed, e.g., items with a correct answer that is never selected by learners with a higher level of knowledge. These items should be removed or rewritten as they fail to measure the learner's knowledge.

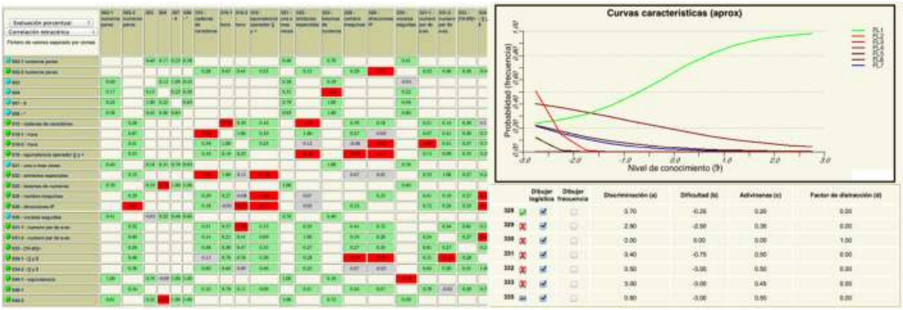
Figure 10 shows a detailed view of the learner model (LM) according to the assessments taken. The intensity of the colour depends on the statistical confidence of the measure. The student model graphical presentation of the LM is delegated to the *Ingrid* system (Conejo et al. 2012). Teachers have access to individual and aggregated models and can compare a student's results to the class average, etc.

If further analyses are needed, Siette can export the results to CSV and ARFF file formats to be analysed with a spreadsheet or with Weka.

Integration Features and Interoperability

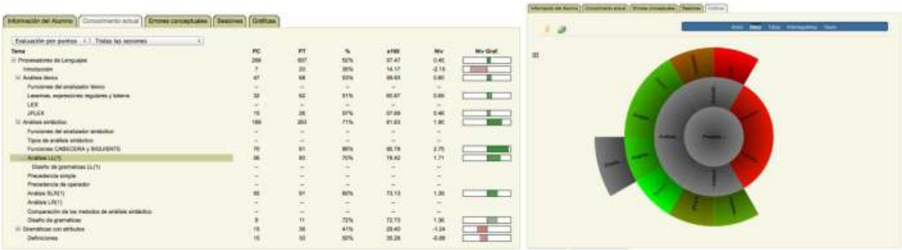
As mentioned, Siette provides a set of web services for system integration. These web services allow secure user creation and login, test delivery, and result exchange. The web services protocol is based on a single sign-on approach, which uses a RSA public/private key schema to secure the communication. In addition to these web services, we have constructed an API for system integration in Java and PHP. There is also a Moodle plug-in that allows integration between Siette and Moodle (supported in Moodle version 1.7 to the current version 2.5). Users created in Moodle have direct access to Siette, while preserving the role they have in Moodle (see Fig. 3).

Siette defines its own XML format for representing the domain concept hierarchy, items, tests, and test session results for backup and restore. This format is called SQTI (Siette Questions and Test Interoperability) and can be used for interoperability with other systems, although currently no other system supports all the features of Siette, as far as we know. Siette can also import items in the Moodle GIFT format and can


Evaluación porcentual
Correlación tetracórica
Curvas caracteristicas (aprox)
Probabilidad (frecuencia)
Nivel de conocimiento (θ)
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0

		Dibujar logística	Dibujar frecuencia	Discriminación (a)	Dificultad (b)	Adivinanza (c)	Factor de distracción (d)
328	✓	☑	□	3.70	-0.26	0.20	0.00
329	☒	☑	□	2.90	-2.50	0.38	0.00
330	☒	☑	□	0.00	0.00	0.00	1.00
331	☒	☑	□	0.40	-0.75	0.50	0.00
332	☒	☑	□	3.50	-3.00	0.56	0.00
333	☒	☑	□	3.00	-3.00	0.45	0.00
335	—	☑	□	0.80	-3.00	0.56	0.00

Fig. 9 Item analysis tools



Evaluación por puntos | Todas las sesiones

Tema	PC	PT	%	a100	Niv	Niv Graf
Procesadores de Lenguajes	288	557	52%	57.47	0.42	
Introducción	7	20	35%	14.17	-2.16	
Análisis léxico	47	58	83%	88.83	0.60	
Funciones del analizador léxico	–	–	–	–	–	
Lexemas, expresiones regulares y tokens	32	62	51%	60.87	0.89	
LEX	–	–	–	–	–	
JFLEX	15	26	57%	57.69	0.46	
Análisis sintáctico	189	263	71%	81.03	1.80	
Funciones del analizador sintáctico	–	–	–	–	–	
Tipos de análisis sintáctico	–	–	–	–	–	
Funciones CABECERA y SIGUIENTE	75	81	88%	95.78	2.75	
Análisis LL(1)	56	80	70%	78.42	1.71	
Diseño de gramáticas LL(1)	–	–	–	–	–	
Precedencia simple	–	–	–	–	–	
Precedencia de operador	–	–	–	–	–	
Análisis SLR(1)	55	91	60%	73.13	1.36	
Análisis LR(1)	–	–	–	–	–	
Comparación de los metodos de análisis sintáctico	–	–	–	–	–	
Diseño de gramáticas	8	11	72%	72.73	1.36	
Gramáticas con atributos	15	38	41%	29.40	-1.24	
Definiciones	15	30	50%	35.28	-0.88	

Fig. 10 Views of the open learner model

provide direct support for IMS QTI 1.2 question definition using a built-in player. The new version (2.0) is not yet supported, but will be in the near future.

The Use of Siette

There is little data on the earliest use of Siette. The current database contains information from 2002. There are 36,000 registered students and around 1000 teachers at the university, who have created a total of 30,000 items. The following tables and figures show how the use of Siette has evolved between 2002 and 2014.

Figure 11 shows the number of active users (i.e. students who have taken at least one assessment session that year). Figure 12 shows the total number of test sessions by access mode. Most users interact with Siette through its main interface (62 % in 2014). In 2007, Moodle was installed at Malaga University Virtual Campus and Siette was integrated as a Moodle activity. The system became far more visible to the university staff. However, most teachers use Moodle questionnaires instead of Siette questionnaires either because they do not know Siette, do not have the time to learn a new tool, or simply because the simpler Moodle assessment questionnaires are sufficient for their needs. The other users have interacted with Siette through other routes, such as the ActiveMath system. Figures 13 and 14 respectively show the number of active teachers (i.e. teachers who have created at least one item) and the number of items created per year.

Most of the content (91 %) of Siette is written in Spanish (see Fig. 15) and was developed at Malaga University, the Polytechnic University of Madrid, and UNED (Spanish Distance Education University). MCQ with single answer is the most popular, followed by open short answer questions (Fig. 16).

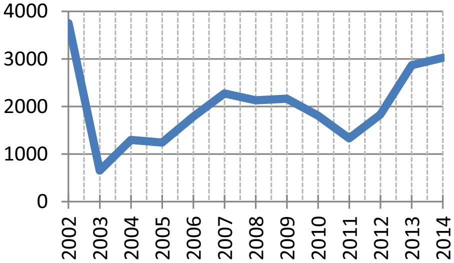


Fig. 11 Number of active students by year

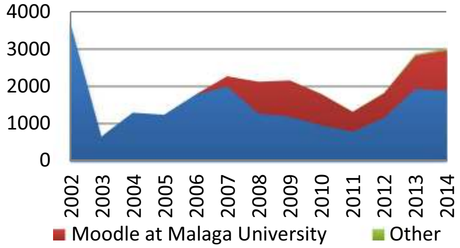


Fig. 12 Number of test sessions by access mode

Some of the most innovative items and particular features of Siette are rarely used (see Fig. 17). Several different hypotheses could explain this fact. Firstly, some users are not comfortable with innovation. It requires a learning effort and the users do not always have sufficient time or interest to improve what they have always been doing in a certain way. Secondly, it requires considerable effort to understand IRT and CAT and it is not easy to explain the results of these to students. It is difficult for students and teachers to accept that the same exam may contain different questions for different people, even if the underlying theory demonstrates the validity of this approach. In Spain, there is a tradition of using the MCQ format and scoring the exams based on the percentage of correct answers; in practice, this situation is very difficult to change. Another obstacle to the use of IRT is the need to previously calibrate the item pool. It is impossible to immediately obtain test results without prior calibration. According to the psychometrics literature, the minimum number of examinees needed for calibration ranges between 100 and 1000, depending on the model used. Classical test theory is applicable to most topics studied by only 30–40 students a year that have large item pools.

Figures 18 and 19 respectively show the evolution of the number of test sessions and questions answered between 2002 and 2014. The figures show that there is a correlation between them. Most of current tests in Siette use the percentage of correct answers as the assessment criterion (642 tests). The application of this assessment criterion is followed by the alternative of item scoring (581 tests). In contrast, IRT was used in only 18 tests.

Finally, random item selection (Fig. 20) was the most popular method used by teachers (1078 tests), followed by fixed order (101 tests), and weighted random (30 tests). This distribution slightly changes depending on the number of test sessions taken.

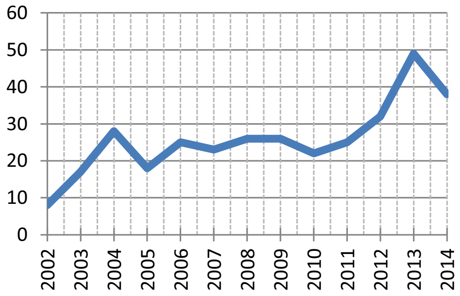


Fig. 13 Number of active teachers by year

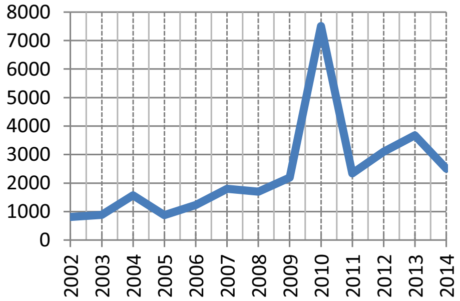


Fig. 14 Number of items created by year

Related Work

Other researchers have addressed problems similar to those encountered during the development and implementation of Siette, some of which have been mentioned above. We now briefly summarize the similarities to and differences between Siette and their approaches.

Knowledge Tracing (KT) (Corbett and Anderson 1994) is a technique used to model student knowledge and learning over time and has mainly been included in cognitive tutors (CT). Knowledge tracing is based on the estimation of four parameters associated with students. Learning Factor Analysis and Performance Factor Analysis (Cen et al. 2006; Pavlik et al. 2009) are modifications of KT that extend it with adaptation capabilities similar to those of CAT. Multidimensional IRT models have also been used in the area of CT and KT (Cen et al. 2008). Pardos and Heffernan (2011) have extended the standard KT model to take into account several item-related difficulties similar to those found in IRT. Bayesian Knowledge Tracing (BKT) has also been applied to CBM (Mitrovic et al. 2003), similar to the manner in which we have attempted to combine IRT and CBM (Gálvez et al. 2013).

Evidence Centered Design (ECD) evolved within the field of psychometrics to assess high-order cognitive skills (Mislevy et al. 2003). The main idea underlying ECD is to separate the evidence model from the diagnostic model and to define different layers in the assessment system.. The main difference between them is that ECD is simply a framework that has to be implemented for each application, whereas Siette is a system that provides functionality that is already implemented for many

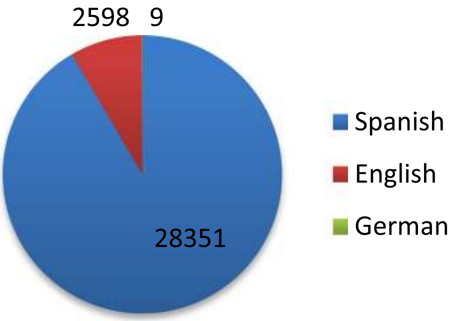


Fig. 15 Number of items by language

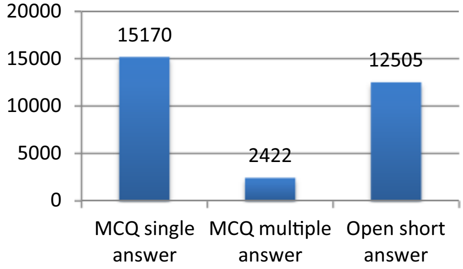


Fig. 16 Number of questions per type

applications. The functionality in SIETTE may help make ECD-style arguments more explicit. For example, the ACED system (Shute et al. 2007) was designed as an assessment for learning system, and many of its features could have been directly implemented in Siette by reusing Siette's components. However, the scope of ECD is broader than that of Siette, which is limited to its implemented features.

Assistment (Razzaq et al. 2005) is another interesting system that explores the possibilities of assessment for learning. It strongly relies on the use of hints and feedback. Assistment and Siette share some common objectives, but Assistment uses sequences of hints and some other features that are not implemented in Siette. On the other hand, Siette includes many features not yet implemented in Assistment and also supports IRT models.

As a general purpose and domain-independent assessment environment, Siette has something in common with LON-CAPA, Moodle, OpenEdX, and other learning management systems (LMS) that include assessment components. A deeper comparison of Siette and each of these systems is beyond the scope of this paper. In general, Siette includes almost all of the features of these assessment components, (e.g., Siette can import any Moodle set of questions defined in a Moodle GIFT file), but the opposite is not the case. IRT and CAT are not present in these LMSs, and so does the hierarchical domain structure and student model, the complex item types, the

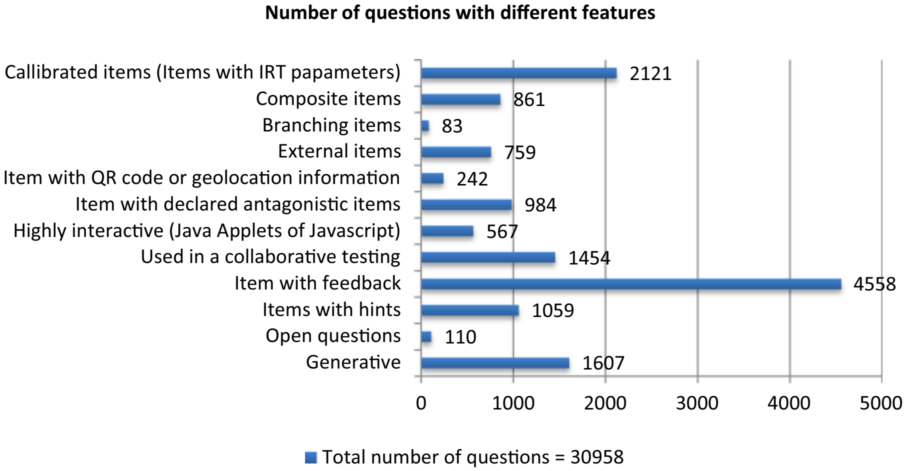


Fig. 17 Number of questions with different features (note that these groups are not exclusive)

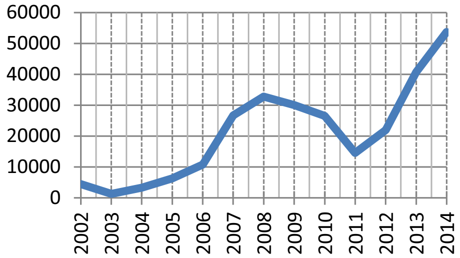


Fig. 18 Number of test sessions per year

collaborative testing features, the pervasive learning assessment facilities, etc. Furthermore, these assessment components of existing LMSs do not provide services that allow their integration into an ITS.

Many other systems share the goal of Siette of adapting question selection to the users' knowledge (Barla et al. 2010; Hsiao et al. 2010). Other systems share the goal of development assessment models based on IRT and other probabilistic assessment algorithms for complex domains. They produce fine-grain representations of students' knowledge (Desmarais et al. 2006; Falmagne and Doignon 2011). Common goals have been semantic integration of adaptive assessment as a component of an ITS and interoperability with open student models (Zapata-Rivera et al. 2007; Sosnovsky et al. 2009). Many other features of Siette, such as collaborative testing (Robinson et al. 2008), spaced repetition (Pavlik and Anderson 2003), or mobile applications and location aware testing (Romero et al. 2009; Santos et al. 2011) have also been another common focus of interest. The most distinctive character of Siette is probably the integration of all these features into a common fully implemented and operative system.

Conclusions

Over the last 16 years, the Siette system has evolved from being a testing tool based on MCQs to a complete automatic assessment environment that supports different assessment methods and that can be used for the formative and summative assessment of multiple types of skills. Different assessment elements and strategies have been explored and integrated within the Siette environment. The advantage of this approach is that the different features can be freely combined in the same assessment session. For example, hints and feedback can be used in pervasive assessment or in a collaborative test, and adaptive models can be used together with misconceptions. The Siette

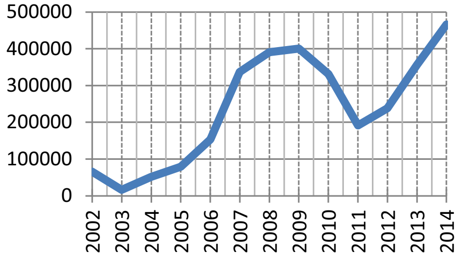


Fig. 19 Number of questions answered per year

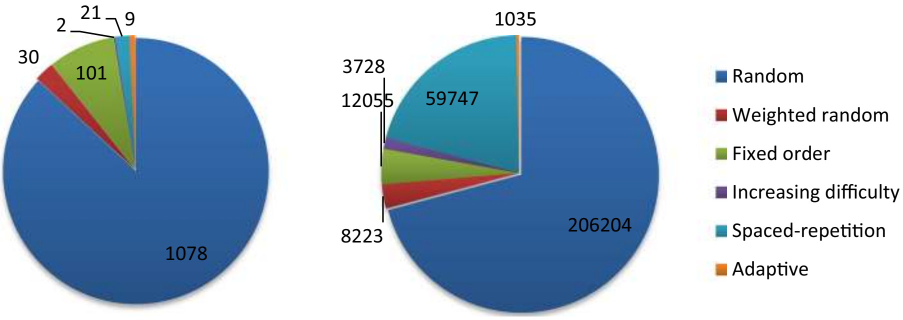


Fig. 20 Number of tests and test sessions by item selection criteria

environment includes three types of learning analytics features: the analysis of student results, the analysis of test validity and reliability, and item analysis.

We consider that the Siette system represents a small but significant contribution to the field of AIED. Firstly, its development has contributed to research on the relationship between psychometric techniques and the assessment problems faced in intelligent tutoring. It is not unusual to find other current research in this field that takes into account IRT and CAT. Secondly, the Siette system addresses the issue of interoperability. Current trends in software development are focused on component reuse. Independent but integrable modules are needed to reduce the cost of ITS development. Last, but not least, Siette has contributed to the exploration of different assessment types and conditions, and has attempted to accommodate them to well-founded assessment models that are driven by collected data and not only by heuristics.

Siette has been mainly used in higher education due to the fact that it was developed in a university environment and tailored to the needs of assessment at this level of education. The application domain includes computer science and engineering, physics, chemistry, maths, biology, botany, economics, and language learning. Some promising results have been obtained by its application to high-school education.

From a practical point of view, Siette has been moderately successful. The number of Siette users has increased every year and not only at Malaga University. The integration of Moodle and Siette has made the system more accessible, which has increased the number of users. However, most of the innovative features of Siette are rarely used, as the data presented in the previous section shows. There are many possible explanations for this situation. On the one hand, many users are unaware of most of the features included in Siette. Effort should be made to document and disseminate the research results among users. In fact, training sessions have been recently been organized for teachers at Malaga University to introduce the features of Siette. On the other hand, the adoption of some of the innovations requires extra effort by teachers. For instance, adaptive behaviour based on IRT requires the teacher to understand this theory, whereas the classic percentage or scoring procedures are well-known and easier to understand by teachers and students. Rich content elicitation, the addition of hints and feedback, misconceptions, and other features also require extra authoring work. This includes maintaining the item bank, analysing and calibrating items, and removing incorrect responses. These challenges could be overcome if the system encapsulates and automatizes some processes to reduce the cognitive load of the teacher.

These problems are common to many intelligent systems, which are commonly used not only by their creators but by many others. Extra effort should be made to provide features such as authoring, analytics, and reporting tools, documentation, and user support channels, in the attempt to persuade individuals to invest time in learning to use the system. A key point is that the final users (teachers) should be able to access the system in a simple way at the beginning and discover advances features as needed. This issue has guided the development of Siette and explains the relatively limited use of its advanced features. However, we consider that this strategy is the correct one to follow if we want intelligent educational systems to have a real impact on current learning practice.

Siette can be freely accessed at www.siette.org. Documentation and further information can be obtained following the links on the first page. Although Siette is not currently an open-source project, free academic use can be granted upon request.

Acknowledgments The authors of this article would like to thank to all the teachers and students who participated in the experiments. This study has been partially funded by the Consejeria de Economía Innovación y Ciencia of the Junta de Andalucía (Spanish regional government) through the research projects TIADA (P07-TIC-03243), PATIO (P08-TIC-4273), and DEDALO (P09-TIC-5105).

References

Arroyo, I., Conejo, R., Guzmán, E., & Woolf, B. P. (2001). An adaptive web-based component for cognitive ability estimation. In J. D. Moore, C. Luckhardt-Redfield, & W. Lewis Johnson (Eds.), *Artificial intelligent in education: AI-ED in the wired and wireless future* (pp. 456–466). Amsterdam: Ios Press.

Barla, M., Bieliková, M., Ezzeddinne, A. B., Kramár, T., Šimko, M., & Vozár, O. (2010). On the impact of adaptive test question selection for learning efficiency. *Computers & Education, 55*(2), 846–857.

Barros, B., Conejo, R., & Guzman, E. (2007). Measuring the effect of collaboration in an assessment environment. In R. Luckin, K. R. Koedinger, & J. Greer (Eds.), *Artificial intelligence in education: Building technology rich learning contexts that work* (Vol. 158, pp. 375–382). Amsterdam: Ios Press.

Cen, H., Koedinger, K. R., & Junker, B. (2006). Learning factors analysis—A general method for cognitive model evaluation and improvement. In M. Ikeda, K. Ashley, & T. Chan (Eds.), *Intelligent Tutoring Systems 8th International Conference* (pp. 164–175). Berlin: Springer.

Cen, H., Koedinger, K. R., & Junker, B. (2008). Comparing two IRT models for conjunctive skills. In B. Woolf, E. Aimer & R. Nkambou (Eds.), *Proceedings of the Proceedings of the 9th International Conference on Intelligent Tutoring Systems*. Montreal, Canada.

Conejo, R., Millán, E., Pérez-de-la-Cruz, J. L., & Trella, M. (2000). An empirical approach to on-line learning in SIETTE. In G. Gauthier, K. VanLehn & C. Frasson (Eds.), *ITS 2000*, LNCS (Vol. 1839, pp. 605–614). Heidelberg: Springer.

Conejo, R., Guzmán, E., Millán, E., Trella, M., Pérez-De-La-Cruz, J. L., & Ríos, A. (2004). *SIETTE: A web-based tool for adaptive testing. International Journal of Artificial Intelligence in Education, 14(1), 29–61*. Amsterdam: Ios Press.

Conejo, R., Guzmán, E., Pérez-de-la-Cruz, J. L., & Millán, E. (2005). Introducing adaptive assistance in adaptive testing. In C.-K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.), *Artificial Intelligence in Education (AIED 2005): Supporting learning through intelligent and socially informed technology* (pp. 777–779). Amsterdam: Ios Press.

Conejo, R., Guzmán, E., Pérez-de-la-Cruz, J. L., & Millán, E. (2006). An empirical study about calibration of adaptive hints in web-based adaptive testing environment. In V. Wade, H. Ashman, & B. Smyth (Eds.), *Adaptive hypermedia and adaptive web-based systems. 4th International Conference AH-2006. Lecture Notes in Computer Science 4018* (pp. 71–80). Berlin: Springer.

Conejo, R., Barros, B., Guzmán, E., & Gálvez, J. (2008). Formative evaluation of the SIETTE collaborative testing environment. In T.-W. Chan, et al. (Eds.), *Proceedings of the 16th International Conference on Computer in Education*, ICCE08 (pp. 297–301). Taipei, Taiwan.

Conejo, R., Barros, B., Guzmán, E., & Gálvez, J. (2009a). Collaborative assessment with SIETTE. In V. Dimitrova, R. Mizogouchi, & B. du Boulay (Eds.), *Artificial Intelligence in Education (AIED-2009)—Building learning systems that care: From knowledge representation to affective modelling* (p. 799). Amsterdam: Ios Press.

Conejo, R., Barros, B., Guzmán, E., & Gálvez, F. (2009b). An experiment to measure learning in a collaborative assessment environment. In V. Dimitrova, R. Mizogouchi & B. du Boulay (Eds.), *Artificial Intelligence in Education (AIED-2009)—Building learning systems that care: From knowledge representation to affective modelling*. Amsterdam, pp. 620–623.

Conejo, R., Trella, M., Cruces, I., & Garcia, R. (2012). INGRID: A web service tool for hierarchical open learner model visualization. In: L. Ardissono & T. Kuflik (Eds.), *UMAP 2011 Workshops*, LNCS, (Vol. 7138, pp. 406–409). Heidelberg: Springer.

Conejo, R., Barros, B., Guzmán, E., & Garcia-Viñas, J. I. (2013). A web based collaborative testing environment. *Computers & Education, 68*, 440–457.

Conejo, R., Garcia-Viñas, J. I., Gaston A., Barros, B. (2015) Technology Enhanced Formative Assessment of Plant Identification. Journal of Science Education and Technology. http://link.springer.com/article/10.1007/s10956-015-9586-0?wt_mc=internal.event.1.SEM.ArticleAuthorOnlineFirst

Conejo, R., Guzmán, E., Perez-De-La-Cruz, J. L., & Barros, B. (2014). An empirical study on the quantitative notion of task difficulty. *Expert Systems with Applications, 41*(2), 594–606.

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*(4), 253–278.

Desmarais, M. C., Meshkinfam, P., & Gagnon, M. (2006). Learned student models with item to item knowledge structures. *User Modeling and User-Adapted Interaction, 16*(5), 403–434.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah: Lawrence Erlbaum Associates Publishers.

Falmagne, J. C., & Doignon, J. P. (2011). *Learning spaces. Interdisciplinary applied mathematics*. Berlin: Springer.

Gálvez, J. (2009). A probabilistic model for student knowledge diagnosis in learning environments. In V. Dimitrova, R. Mizogouchi & B. du Boulay (Eds.), *Artificial Intelligence in Education (AIED-2009)—Building learning systems that care: From knowledge representation to affective modelling* (Vol 200. pp. 759–760). IOS Press.

Gálvez, J. (2012). Modelado Probabilístico del Alumno en Entornos Inteligentes de Resolución de Problemas Educativos. Doctoral Dissertation. (In Spanish).

Gálvez, J., Guzmán, E., & Conejo, R. (2009a). A blended E-learning experience in a course of object oriented programming fundamentals. *Knowledge-Based Systems, 22*(4), 279–286.

Gálvez, J., Guzmán, E., Conejo, R., & Millán, E. (2009b). Student knowledge diagnosis using item response theory and constraint-based modeling. In V. Dimitrova, R. Mizogouchi & B. du Boulay (Eds.), *Artificial Intelligence in Education (AIED-2009)—Building learning systems that care: from knowledge representation to affective modelling* (Vol. 200, pp. 291–299). IOS Press.

Gálvez, J., Guzmán, E., & Conejo, R. (2009c). Data-driven student knowledge assessment through ill-defined procedural tasks. In P. Meseguer, L. Mandow & R. M. Gasca (Eds.), *CAEPIA 2009 Selected papers*. LNCS(LNAI) (Vol. 5988, pp. 233–241). Heidelberg: Springer.

Gálvez, J., Guzmán, E., & Conejo, R. (2010). Using intelligent adaptive assessment models for teaching mathematics. In J. L. Galán García, G. Aguilera Venegas & P. Rodríguez Cielos (Eds.), *Book of Abstracts of Technology and its Integration into Mathematics Education (TIME 2010)* (p. 108). Málaga (Spain).

Gálvez, J., Guzmán, E., & Conejo, R. (2012). Exploring quality of constraints for assessment in problem solving environments. In S. A. Cerri, W. J. Clancey, G. Papadourakis & K. Panourgia (Eds.), *Proceedings of the 11th International Conference on Intelligent Tutoring Systems (ITS 2012)* (pp. 310–319) Chania (Greece), June 2012.

Gálvez, J., Conejo, R., & Guzmán, E. (2013). Statistical techniques to explore the quality of constraints in constraint-based modeling environments. *International Journal of Artificial Intelligence in Education, 23*, 22–49.

Greer, J., & McCalla, G. (Eds.) (1994). *Student modeling: The key to individualized knowledge-based instruction*. NATO ASI Series F (Vol. 125). Berlin: Springer-Verlag.

Guzmán, E. (2005). Un Modelo de Evaluación Cognitiva basado en Tests Adaptativos Informatizados para el diagnostic en Sistemas Tutores Inteligentes. Doctoral Dissertation. (In Spanish).

Guzmán, E., & Conejo, R. (2002). Simultaneous evaluation of multiple topics in Siette. In S. Cerri, G. Gouardères, & F. Paraguaçu (Eds.), *Intelligent tutorial systems, 6th international conference. Lecture notes in computer science 2363* (pp. 739–748). Berlin: Springer.

Guzmán, E., & Conejo, R. (2004a). A library of templates for exercise construction in an adaptive assessment system. *Technology, Instruction, Cognition and Learning, 2*, 21–60.

Guzmán, E., & Conejo, R. (2004b). A model for student knowledge diagnosis through adaptive testing. In J. Lester, R. M. Vicari & F. Paraguaçu (Eds.), *Intelligent tutoring systems*. 7th International Conference, ITS 2004. Lecture Notes in Computer Science 3220 (pp. 12–21). Berlin: Springer.

Guzmán, E., & Conejo, R. (2004c). A brief introduction to the new architecture of Siette. In P. De Bra & W Nejdl (Eds.), *Adaptive hypermedia and adaptive web-based systems*. Lecture Notes in Computer Science 3137 (pp. 405–408). Berlin: Springer.

Guzmán, E., & Conejo, R. (2005a). Towards efficient item calibration in adaptive testing. In L. Ardisono, P. Brna & A. Mitrovic (Eds.), *User modelling 2005. 10th International Conference UM-2005*, Lecture Notes in Artificial Intelligence 3538 (pp. 402–406). Berlin: Springer.

Guzmán, E., & Conejo, R. (2005b). Self-assessment in a feasible, adaptive web-based testing system. *IEEE Transactions on Education, 48*(4), 688–695.

Guzmán, E., & Conejo, R. (2015). Measuring misconceptions through item response theory. In C. Conati, N. Heffernan, A. Mitrovic & M. F. Verdejo (Eds.), *Artificial Intelligence in Education, (AIED-2015)*, LNCS (Vol. 9112, pp. 608–611). Springer International Publishing.

Guzmán, E., Conejo, R., Hontangas, P., Olea, J., & Ponsoda, V. (2000). A comparative study of IRT and classical item parameter estimates web-based and conventional test administration. In *International Test Commission's Conference on Computer-Based Testing and the Internet*. Winchester (England).

Guzmán, E., Conejo, R., & García-Hervás, E. (2005). An authoring environment for adaptive testing. *Educational Technology & Society, 8*(3), 66–76.

Guzmán, E., Conejo, R., & Pérez-de-la-Cruz, J. L. (2007a). Adaptive testing for hierarchical student models. *User Modeling and User-Adapted Interaction, 17*(1–2), 119–157.

Guzmán, E., Conejo, R., & Pérez-de-la-Cruz, J. L. (2007b). Improving student performance using self-assessment tests. *IEEE Intelligent Systems, 22*, 46–52.

Guzmán, E., Conejo, R., & Gálvez, J. (2010). A data-driven technique for misconception elicitation. In *Proceedings of the International Conference on User Modelling, Adaptation and Presentation UMAP-2010, June, Big Island of Hawaii (USA)*, LNCS (Vol. 6075, pp. 243–254). Berlin: Springer.

Hernando, M. (2011). Student procedural knowledge inference through item response theory. In J. A. Konstan, R. Conejo, J. L. Marzo & N. Oliver (Eds.), *User Modelling, Adaption and Personalization, (UMAP-2011)*, LNCS (Vol. 6787, pp. 426–429). Springer International Publishing.

Hernando, M., Guzmán, E., & Conejo, R. (2013a). Measuring procedural knowledge in problem solving environments with item response theory. In H. Lane, K. Yacef, J. Mostow & P. Pavlik (Eds.), *Artificial Intelligence in Education Artificial Intelligence in Education* (AIED-2013), LNCS. (Vol. 7926, pp. 653–656). Berlin: Springer International Publishing.

Hernando, M., Guzmán, E., & Conejo, R. (2013b). Validating item response theory models in simulated environments. In G. McCalla & J. Champaign, H. Lane, K. Yacef, J. Mostow & P. Pavlik (Eds.), *AIED 2013 Simulated Learners Workshop. Artificial Intelligence in Education*, LNCS (Vol. 7926, pp. 954–955). Berlin: Springer International Publishing.

Hsiao, I. H., Sosnovsky, S., & Brusilovsky, P. (2010). Guiding students to the right questions: adaptive nnavigation support in an e-Learning system for Java programming. *Journal of Computer Assisted Learning, 26*(4), 270–283.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Oxford: Addison-Wesley.

McCalla, G. I., & Greer, J. E. (1994). Granularity-based reasoning and belief revision in student models. In J. Greer & G. McCalla (Eds.), *Student modeling: The key to individualized knowledge-based instruction*, NATO ASI Series F (Vol. 125). Berlin: Springer

Melis, E., Andres, E., Budenbender, J., Frischauf, A., Goduadze, G., Libbrecht, P., & Ullrich, C. (2001). ActiveMath: a generic and adaptive web-based learning environment. *International Journal of Artificial Intelligence in Education (IJAIED), 12*, 385–407.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Mahwah: Lawrence Erlbaum Associates Publishers.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessment. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3–62.

Mitrovic, A. (2012). Fifteen years of constraint-based tutors: what we have achieved and where we are going. *User Modeling and User-Adapted Interaction, 22*(1-2), 39–72.

Mitrovic, A., Koedinger, K. R., & Martin, B. (2003). A comparative analysis of cognitive tutoring and constraint-based modeling. In P. Brusilovsky, et al. (Eds.), *Proceedings of the 9th International Conference on User Modeling (UM2003)*, LNAI 2702 (pp. 313–322). Berlin: Springer.

Pardos, Z. A., Heffernan, N. T. (2011). KT-IDEM: Introducing item difficulty to the knowledge tracing model. In J. Konstan, R. Conejo, J. L. Marzo & N. Oliver (Eds.), *Proceedings of the 19th International Conference on User Modeling, Adaptation and Personalization*, Lecture Notes in Computer Science (Vol. 6787, pp. 243–254)

Pavlik, P. I., & Anderson, J. R. (2003). An ACT-R model of the spacing effect. In F. Detje, D. Dorner, & H. Schaub (Eds.), *Proceedings of the Fifth International Conference of Cognitive Modelling* (pp. 177–182). Germany: Universitats-Verlag Bamberg.

Pavlik, P. I., Cen, H., & Koedinger, K. R. (2009). Performance factors analysis—A new alternative to knowledge tracing. In V. Dimitrova, R. Mizogouchi & B. du Boulay (Eds.), *Artificial Intelligence in Education (AIED-2009)—Building learning systems that care: From knowledge representation to affective modelling* (Vol. 200, pp. 531–538). IOS Press.

Pérez-de-la-Cruz, J. L., Conejo, R., & Guzmán, E. (2005). Qualitative and quantitative student models. In C.-K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.), *Artificial Intelligence in Education (AIED-2005): Supporting learning through intelligent and socially informed technology* (pp. 531–538). Amsterdam: Ios Press.

Polson, M. C., & Richardson, J. J. (Eds.). (2013). *Foundations of intelligent tutoring systems*. London: Psychology Press.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*(4), 611–630.

Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N. T., Koedinger, K. R., Junker, B., & Rasmussen, K. P. (2005). The assistment project: Blending assessment and assisting. In C.-K. Looi, G. McCalla, B. Bredeweg, & J. Breuker (Eds.), *Artificial Intelligence in Education (AIED-2005): Supporting learning through intelligent and socially informed technology* (pp. 555–562). Amsterdam: Ios Press.

Rios, A. (1998) Siette: Sistema de Evaluación de Tests para la TeleEducación, Master thesis. Department of Computer Science, University of Malaga, Spain. (In Spanish).

Rios, A., Pérez de la Cruz, J. L., & Conejo, R. (1998). Siette: Intelligent evaluation system using tests for TeleEducation. In Proc. of Workshop "WWW-Based Tutoring" at 4th International Conference on Intelligent Tutoring Systems, San Antonio, TX.

Rios, A., Millán E., Trella, M., Pérez de la Cruz, J. L., & Conejo, R. (1999). Internet based evaluation system. In S. Lajoie & M. Vivet (Eds.), *Artificial intelligence in education, open learning environments: New computational technologies to support learning, exploration and collaboration*, AIED-99 Le Mans, France (pp. 387–394). IOS Press.

Robinson, D. H., Sweet, M., & Mayrath, M. (2008). A computer-based, team-based testing system. In D. H. Robinson, J. M. Royer & G. Schraw (Series Eds.), *Recent innovations in educational technology that facilitate student learning current perspectives on cognition, learning and instruction* (pp. 277–290).

Romero, C., Ventura, S., & De Bra, P. (2009). Using mobile and web-based computerized tests to evaluate university students. *Computer Applications in Engineering Education, 17*(4), 435–447.

Santos, P., Pérez-Sanagustín, M., Hernández-Leo, D., & Blat, J. (2011). QuesTInSitu: from tests to routes for assessment in situ activities. *Computers & Education, 57*(4), 2517–2534.

Shute, V. J., Hansen, E. G., & Almond, R. (2007). Evaluating ACED: The impact of feedback and adaptivity on learning. In *Frontiers in artificial intelligence and applications* (pp. 158, 230). IOS Press.

Sosnovsky, S., Brusilovsky, P., Yudelson, M., Mitrovic, A., Mathews, M., & Kumar, A. (2009). Semantic integration of adaptive educational systems. In *Advances in ubiquitous user modelling* (pp. 134–158). Berlin: Springer.

Trella, M., Conejo, R., & Guzmán, E. (2000). A web-based socratic tutor for trees recognition. In P. Brusilovsky, O. Stock & C. Strapparava (Eds.), *Adaptive hypermedia and adaptive web-based systems AH-2000*, LNCS (Vol. 1892. pp. 239–249). Springer.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. New York: Routledge.

Zapata-Rivera, D., Hansen, E., Shute, V. J., Underwood, J. S., & Bauer, M. (2007). Evidence-based approach to interacting with open student models. *International Journal of Artificial Intelligence in Education, 17*(3), 273–303.