Expert Systems with Applications 41 (2014) 594–606

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

An empirical study on the quantitative notion of task difficulty

Ricardo Conejo, Eduardo Guzmán, Jose-Luis Perez-de-la-Cruz, Beatriz Barros *

E.T.S. Ingeniería Informática, Universidad de Malaga, 29071 Malaga, Spain

ARTICLE INFO

Keywords:
Web-based educational system
Intelligent tutoring systems
Knowledge assessment
IRT
CAT
Item difficulty
Item calibration
Difficulty estimation

ABSTRACT

Most Adaptive and Intelligent Web-based Educational Systems (AIWBES) use tasks in order to collect evidence for inferring knowledge states and adapt the learning process appropriately. To this end, it is important to determine the difficulty of tasks posed to the student. In most situations, difficulty values are directly provided by one or more persons. In this paper we explore the relationship between task difficulty estimations made by two different types of individuals, teachers and students, and compare these values with those estimated from experimental data. We have performed three different experiments with three different real student samples. All these experiments have been done using the SIETTE web-based assessment system. We conclude that heuristic estimation is not always the best solution and claim that automatic estimation should improve the performance of AIWBES.

1. Introduction

The advent of the Internet has entailed the apparition of several kinds of tools. From an educational perspective, the Internet is a repository of information that both teachers and students can use to their own benefit. However, the evolution of technologies used to develop web-based tools has led to the use of new and more sophisticated systems. These new tools offer the student a tutored learning process, emulating the behavior of a teacher in a classroom. Such systems are called *Adaptive and Intelligent Web-based Educational Systems* (AIWBES) (Brusilovsky & Peylo, 2003) and they are the evolution of two families of systems: Intelligent Tutoring Systems and *Adaptive Hypermedia Systems* (Brusilovsky, 2001). The first have emerged as a result of applying Artificial Intelligence techniques to *Computer-Assisted Learning* (CAL) Systems. *Intelligent Tutoring Systems* (ITS) are also influenced by two other knowledge areas, like Cognitive Psychology and Educational Research. Initially, they were intended to partially automate the task of providing the student with individualized and self-paced learning instruction.

In AIWBES, the learning process is adapted to student needs. This adaptation requires the elicitation and updating of student models. A *student model*, also called *leaner model*, (LM) represents the perception of the system about the learner (VanLehn, 1988). Selecting the right task or question to pose according to the student model is a central topic in intelligent learning systems. (Barla et al., 2010). The quality of an AIWBES will be determined by the scope and quality of the data stored in the learner model, and by the ability of the system to update this model appropriately. This update is usually carried out on the basis of evidence generated from student examinations. Student responses to tasks are raw data which should be converted into information and used to update learner models. The selection of the most appropriate task and the process of updating learner models depend on the properties of the task. Perhaps one of the most relevant properties is task difficulty. Everybody has a subjective notion of what difficulty means and, if we asked a set of persons to give a precise definition of it, they would surely supply related but different statements.

One of the most used techniques for student knowledge diagnosis is testing. There are well-known psychometric theories that relate observed student responses to his/her knowledge state. Most of the tests we find in AIWBES are based on the Classical Test Theory (CTT). This theory, although easy to apply, does not guarantee reliable and invariant diagnosis. Item Response Theory, (IRT) appeared later to solve some of those problems.

Both theories, i.e. CTT and IRT, provide statistical definitions on the concept of difficulty and use data-driven mechanisms to compute the *difficulty* value. However, we can find that in practice most systems use *estimations provided by human "experts"*. There are also some "mathematical" proposals to estimate the *difficulty* from a set of features of the task, such as its complexity or the number of concepts involved, by means of a formula that predicts the *difficulty* or the student performance.

To sum up, there are three different approaches for estimating the "*difficulty*" of a task:

- *Statistical*, that is, estimating the *difficulty* from a previous sample of students.
- *Heuristic*, that is, by human "experts" direct estimation.

* Corresponding author. Tel.: +34 952 13356.
E-mail address: bbarros@lcc.uma.es (B. Barros).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.07.084

- *Mathematical*, given a formula that predicts the *difficulty* in terms of the number and type of concepts involved in the task.

Statistical approaches require a previous definition of the concept of *difficulty*. So it is commonly associated with CTT or IRT assessment (see Section 2.1), but there is an increasing interest in the ITS and AIWBES community for data mining methods to adjust and fine tune system performance (Romero & Ventura, 2010).

On the other hand, *heuristic* approaches are common in ITS and AIWBES, (see Section 2.2), but IRT assessment sometime use *heuristic* estimation of the item parameters. Teachers, or course creators, are commonly the "experts" that estimate the *difficulty* but there are some experience of using the students as "experts" (see Section 2.3).

What we have called *mathematical* approach can also be viewed as a complex form of heuristic, because the formula itself and the parameters involved are also given by human experts. This approach is mainly used in ITS and AIWBES (see Section 2.2), but also in IRT assessment, for instance to predict the parameters of an item generated from a template (Geerlings, van der Linden, & Glas, 2013). However, *mathematical* approaches are commonly related to complex tasks or problems. In this paper we will focus on simple tasks, like test questions and compare the statistical and heuristic approaches to the *difficulty* parameter estimation.

Another dimension of the problem is time. Parameters need to be configured in some way before the system can be used. If difficulty parameters are estimated *heuristically* they mostly remain unchanged forever because the estimation requires a high costly human effort. On the other hand, there is a cold start problem for the *statistical* approach. This is the case of some IRT models, that require hundred of data to calibrate. Mixed approaches have been used in practice, like a heuristic initial estimation followed by a statistical updating (see Section 2.2). Other authors propose heuristic formulas to continuously update *difficulty* values, based on methods like the Elo rating (Klinkenberg, Straatemeier, & van der Maas, 2011).

This paper tries to contribute to some open research questions: Do statistical and heuristic estimations of difficulty correlate? Are heuristic estimations consistent? Do teachers' estimations and students' estimations correlate? Are heuristic estimations always reliable?

In this work we have carried out several experiments in order to study whether human expert (teacher/student) estimations are similar to *difficulty* values inferred by applying data-driven techniques. We have also explored the alignment of teacher and student viewpoints regarding the quantitative notion of task *difficulty*. Our aim is to focus on the relevance of having a clear understanding of what task *difficulty* represents, especially in AIWBES where educational instruction is adapted to the student needs.

In the next section, primary devoted to the background of this research, we introduce some notions about student modeling and knowledge diagnosis. Test theories and how they define the *difficulty* are considered. We also review some intelligent educational systems, focusing especially on how they manage the task *difficulty*. Section 3 introduces the SIETTE system, which has been used as a workbench to support these experiments. Section 4 describes three different experiments performed with real students and shows and discuss the obtained results. Finally, in Section 5 our results are summarized and some conclusions are drawn.

2. Theoretical background and related work

In this section we present some theoretical background related to the work presented in this paper and analyze different formal and informal definitions of the concept of *difficulty*. As we will see, it is closely related to the problem of knowledge diagnosis. Elements used for knowledge diagnostic purposes are generically called tasks. Tasks are the most interactive part of an assessment, and their main purpose is to elicit evidences (observables) about proficiencies (unobservables) (Shute, Graf, & Hansen, 2005).

Two main framework will be presented: formal test theories CTT and IRT where tasks are usually simple questions, and where *difficulty* has a clear meaning; and the ITS and AIWBES where the *difficulty* of tasks is defined and used in different ways.

The section continues with a summary of previous work about the estimation of the *difficulty* of assessment tasks either by teachers and/or students, analyzing the alignment between teachers' and students' point of view regarding problem solving complexity and strategies to estimate it. Although this is a very interesting question, we have not found many studies about task *difficulty* estimation. To find relevant studies a wide variety of computerized databases were used including Educational Resources Information Center (ERIC), The ISI Web of Knowledge, ScienceDirect, and Google Scholar. The following keywords were combined: *difficulty level, assessment difficulty, item difficulty, task difficulty, calibration and estimation*. Next, the 'snowball method' was employed and the references in the selected articles for additional works were reviewed, and also those articles that cite the previously found papers.

2.1. Task difficulty and knowledge diagnosis in CTT and IRT

2.1.1. Classical Test Theory (CTT)

CTT was first used at the beginning of the 20th century and has been used ever since. According to this theory, the knowledge (ability or true score) of a student is defined as the expected value obtained by a student in a certain test. Given a student s, who takes a test t, his/her knowledge can be expressed as follows:

$$Y_{st} = \tau_{st} + \varepsilon_{st}$$

where Y_{st} is a random variable representing the observed score of subject s when answering test t. This is also called the test score. It is composed of two parts: the true score (τ_{st}) and the measurement error (ε_{st}). Neither is observable. Y_{st} can be computed from the number of questions answered correctly or any other heuristic. In turn, the true score is a random variable with normal distribution with mean equal to zero and unknown variance.

CTT assumes that true score and error are not correlated. Therefore, if we take two different measurements, the errors we obtain are independent of each other. The error measurement is independent of the true score. In this theory items are characterized by two parameters: the *difficulty*, that is, the portion of students who answered the item successfully, and the *discrimination* factor, whose value is a correlation between the item and the test score.

CTT has several limitations, e.g. the knowledge measurement is strongly linked to test features. This means that when we measure student knowledge, we do not obtain an absolute quantitative measurement of his/her knowledge, but rather a value that depends on the test taken. This makes it very difficult to compare students who have taken different tests. Likewise, item parameters represent features of a certain population, therefore are not generic. As a result, the *difficulty* of an item will strongly depend on the knowledge levels of those individuals whose performance is used to infer the *difficulty* and vice versa.

On the other hand, CTT is easy to apply in several situations (Hambleton & Jones 1993). In addition, unlike other theories such as IRT, this theory has fewer requirements, e.g. it requires fewer examinees. Traditional test-based assessment criteria (percentage of success, score obtained, etc.) are in keeping with this theory.

2.1.2. Item Response Theory

This theory is based on two main principles: (a) Student performance in a test can be explained according to their knowledge level, which can be measured as an unknown numeric value θ. (b) The performance of a student with an estimated knowledge level answering an item*i* can be probabilistically predicted and modeled by means of a function called the *Item Characteristic Curve* (ICC). It expresses the probability that a student with certain knowledge level θ will answer the item correctly. ICCs must be calibrated before being used. In the calibration process, each ICC is statistically determined from datasets of students who have taken a test previously. From these results, calibration can be done.

ICCs can be characterized by means of known functions (parametric models) or taken directly from the statistical results (non-parametric models). In the category of parametric models, there are several functions that characterize ICCs. The most common functions are the family of logistic curves of one, two or three parameters (1PL, 2PL or 3PL) defined as follows:

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

where $u_i = 1$ means that the student has successfully answered item *i*. If the student answers incorrectly, $P(u_i = 0|\theta) = 1 - P(u_i = 1|\theta)$. θ is the student's knowledge level, i.e. what is being measured in the test. Knowledge level θ, takes real values in the interval $[-\infty, \infty]$, but in practical application it is considered only within the interval $[-4.0, 4.0]$ or $[-3.0, 3.0]$. Finally, the three parameters that determine the shape of this curve are:

- *Discrimination factor*(a_i): It is proportional to the slope of the curve. High values indicate that the probability of success from students with a knowledge level higher than the item difficulty is high.
- *Difficulty*(b_i): It corresponds to the knowledge level at which the probability of answering correctly is the same as answering incorrectly. The range of values allowed for this parameter is the same as the ones allowed for the knowledge levels.
- *Guessing factor*(c_i): It is the probability that a student with no knowledge at all will answer the item correctly by randomly selecting a response.

This function is used for modeling the 3PL model. If we assume that *guessing* factor is zero, the function obtained is the 2PL model. If we also assume that *discrimination* factor is always equal to one, then the resulting function is the 1PL model, also called the *Rasch* model.

The models presented above are dichotomous, since they consider two possible responses, i.e. correct or incorrect. We can find also polytomous models, where several answers are possible for each question and each answer has its own characteristic curve. They are more informative, although they make calibration process more difficult because instead of calibrating only one curve per item, we must calibrate one curve per answer per item.

As mentioned before, IRT can be used to determine the student knowledge state. In this theory, the inference process consists of calculating a probability distribution curve $P(\theta u_1, \ldots, u_n)$, where $u_1, \ldots, u_n$ is the vector with the responses the student selects for each test item, and θ is the knowledge in the concept whose value is being estimated. One of the most popular estimation techniques is the Bayesian method. It applies Bayes theorem to calculate student knowledge distribution after taking a test with n items:

$$P(\theta|u_i, \ldots, u_n) \propto \sum_{i=1}^{n} P(u_i|\theta)P(\theta)$$

where $P(\theta)$ represents the a priori student knowledge distribution. Several alternatives may be used to obtain the new assessment of the student knowledge distribution regarding the value of $P(\theta)$. Perhaps the most commonly used is to consider a flat distribution where all knowledge levels have the same probability. Another alternative is to assume a knowledge probability distribution corresponding to the population used for test item calibration. Once the student's knowledge probability distribution is obtained, the knowledge level is usually computed using one of the following mechanisms. The first consists of calculating the mean (or expected value) of the distribution. This strategy is called *Expectation a posterior* (*EAP*). The other alternative is the distribution mode. This method is called *Maximum a posteriori* (*MAP*). There are computer software packages that implement these techniques that are commonly used by the IRT community, like *Bilog-MG* (Zimowski, Muraki, Mislevy, & Bock, 1996), *Multilog* (Thissen, 2003), *Parscale* (Muraki & Bock, 1997), *Testfact* (Wilson, Wood, & Gibbons, 1991), *ICL* (Hanson, 2002), or the *R* package *plink*. (Weeks, 2010).

IRT can also be used to determine the most appropriate item to be posed to the student at each point of the test, and also to decide whether the knowledge estimations are accurate enough to stop posing questions. These two uses, combined with the student knowledge state inference, form part of the *Computerized Adaptive Testing Theory* (CAT) for further information about IRT and CAT we refer the interested reader to classical textbooks. (Hambleton, Swaminathan, & Rogers, 1991; Wainer, Dorans, Flaugher, Green, & Mislevy, 2000).

2.2. Tasks difficulty and knowledge diagnosis in ITS and AIWBES

The main goal of an educational system is that students learn new concepts and, accordingly, that his/her domain knowledge and comprehension should increase. As a consequence, student models must be updated to take into account the changes in his/her knowledge state. However the communication channel between the student and the system, where he/she is being tutored, is very restrictive as the system can only measure knowledge directly by monitoring interaction with the student. The process of inferring student characteristics from the observation of his/her behavior is called student diagnosis (VanLehn, 1988). This diagnosis refers to: (a) all those observable features stored in terms of specific functions; (b) internal features which must be inferred from the information stored and relevant to the learning process; and (c) the method used to extract this information through student monitoring and tracking.

The presence of uncertainty is also an important factor that leads to errors in the diagnostic process. This uncertainty may be a consequence of errors and approximations during the data analysis process or may be a result of the abstract nature of human perception and/or information loss caused by quantification (Grigoriadou, Kornilakis, Papanikolaou, & Magoulas, 2002).

From the Artificial Intelligence point of view, in student knowledge diagnosis, the use of a reliable method is crucial. This method should be able to analyze effectively (in the same way as a teacher would do) measurements of student behavior. From these data, the system should make estimations about his/her performance, updating the student model accordingly. However most systems use diagnostic procedures based on heuristics. Consequently, the results obtained lack credibility. In addition, such systems propose paradigms that are not viable from a practical point of view. In general, their implementations depend on requirements that are difficult to satisfy. Another disadvantage of these kinds of systems is that they are applied to specific domains, and are therefore difficult to extrapolate to other areas.

In the literature there are many systems in which task *difficulty* is interpreted from different perspectives. For instance, OLAE

(VanLehn & Martin, 1997) is an assessment system for learning Newtonian physics. This system uses Bayesian user models and provides different types of tasks such as quantitative problem solving activities (where students must compute physical parameters), or studying activities (where students can observe how a problem can be solved step by step) or qualitative problem solving activities (multiple-choice items). The *difficulty* of these tasks has been previously determined by human experts. Additionally, one of the student activities in this system is to numerically estimate the *difficulty* of quantitative problems and to indicate the factors (from a list) that make the task more (or less) *difficult*. However, this information is not used to update the student model.

MDF (mixed numbers, fractions, and decimals) is a mathematics tutor (Beck, Stern, & Woolf, 1997). This system presents problems of addition, subtraction, multiplication, and division of whole numbers, fractions, mixed numbers, and decimals. Each problem is assigned to the topic it evaluates. Furthermore, each topic is broken down into sub-skills corresponding to the steps of the problem solving process. The problem difficulty was estimated a priori following the philosophy; the more sub-skills required to solve a problem, the harder the problem. In addition, other issues are considered when estimating this difficulty, such as the area being assessed.

SQL-Tutor (Mayo & Mitrovic, 2000) is an adaptive and knowledge-based teaching system that supports students learning SQL using a set of problems, which require solving. The adaptation of instruction is done by adjusting the complexity of the problems presented to the students and by generating feedback messages during the process. The *difficulty* of each problem is assigned a priori by an expert in terms of its wording, the constructs needed for its solution, the number of tables/attributes involved, etc.

ELM-ART II (Weber & Brusilovsky, 2001) is an intelligent web-based educational system, designed to support learning LISP programming. It includes tests and exercises. According to their authors, this system was one of the first to include a module for testing within its architecture. In this system, items are grouped and assigned to a knowledge unit. For each item a *difficulty* parameter and a weight are defined and both values are fixed. The weight depends on the group the item belongs to. The *difficulty* determines how much evidence is added to the confidence value of the related concepts when the test item is solved correctly. The confidence value represents the student knowledge state in a unit. After each student's item response in a test, his/her confidence value is updated by multiplying the difficulty and the weight of the item, when the answer is correct. Otherwise, this product is also multiplied with an error factor and subtracted from the confidence factor.

A similar approach is used in QuizGuide (Sosnovsky, Lee, Zadorozhny, & Zhou, 2008), an adaptive assessment for Java programming questions. The system tries to predict the subjective difficulty of a question for each student according to a formula that takes into account the number of concepts involved in each question, their relative weight and the concepts mastered by the student. The weights are defined heuristically by human experts.

ASSISTment (Feng, Heffernan, & Koedinger, 2009) is a web-based math tutoring system for 7th–12th grade students. The system helps the student learn the required knowledge by breaking the problem into sub-questions called scaffolding or by giving the student hints on how to solve the question. The different *difficulty* of the questions has not been considered in the initial design of the system. However authors have recently develop a computer adaptive testing called PLACEments (Whorton, 2013) as an extension of ASSISTment. The system tries to automatically predict the student performance in subsequent questions based on previous responses, using data mining techniques.

Knowledge Tracing (KT) (Corbett & Anderson, 1995) is a technique to model student knowledge and learning over time. It is used to predict the student performance based on the estimation of the probability of having learned the skills involved in the question resolution. It is based on the estimation of four parameters: the initial knowledge, the learn rate and the guess and slip rate. Individual differences are achieved defining a set of weights associated to each student that personalize the four parameters. Parameters and weighs are automatically calibrated from previous students' data. Pardos and Heffernan (2011) have extended the standard KT model to take into account different item difficulty. As they say: "*Models like IRT that take into account item difficulty are strong at prediction, and models such as KT that infer skills are useful for their cognitive diagnostic results*".

IRT has also been used in the field of ITS and AIWBES. The SIETTE system (Conejo et al., 2004) was developed as an independent tool that can be integrated into an ITS. It was integrated in the ActiveMath-1 architecture (Melis et al., 2001) and MEDEA (Trella, Conejo, Guzmàn, & Bueno, 2003). A description of this system is included in Section 3. Barla et al. (2010) describes a similar system that combines IRT with an heuristic selection based on the questions' concept and the history of questions previously posed. SIETTE includes an automatic selection of concepts as an extension of IRT framework based also on item difficulties and the precision of the estimation of the student knowledge level (Guzmán, Conejo, & Pérez-de-la-Cruz, 2007b).

IRT is also the core of the PEL-IRT system (Chen, Lee, & Chen, 2005). The system was later modified to support based on a fuzzy version of IRT (Chen & Duh, 2008). They also evolved from a "*voting approach to determine difficulty parameters of the courseware by integrating experts' decision and learners' voting*" to "*statistic-based methods through a conscientious test process to determine difficulty parameters*". Similar conclusions has been achieved independently by Jeremic, Jovanovic, and Gaševic (2012) "*Diagnostic module uses data about a question's difficulty level and time necessary to solve it provided by a human teacher. . . (we also) performs analysis of the system's logs data, compares this data with question's difficulty level and time necessary to answer the question. We strongly believe that this model could significantly*". Unfortunately, in both cases their paper does not include any experiments that support their decision.

Many other authors have implicitly recognized this problem, and recently the interest for an automatic calibration of difficulty parameters, or automatic adaptation of tasks selection in ITS and AIWBES systems has increased. There are many proposals that try to find out a solution using different AI techniques, like fuzzy logic (Chrysafiadi & Virvou, 2012), Bayesian networks (Millán, Descalço, Castillo, Oliveira, & Diogo, 2013) neural networks (Cabada, Barrón Estrada, & Reyes García, 2011) or genetic algorithms (Verdu, Verdu, Regueras, de Castro, & García 2012).

2.3. Teacher's and students' tasks difficulty estimation

In psychometry there has been always an interest for the estimation of item parameters by experts. In the literature they are also called judges or panellists. Lee (1996) suggested that students could estimate problem difficulty more accurately than teachers. However, teachers' estimation has received more attention than students' estimation. Impara and Plake (1998) conclude that teachers' estimation has high accuracy in the average, but underestimated the performance of the borderline students. Plake and Impara (2001) relate the previous result to a lack of training of teachers in the previous experiments. They found that teacher's estimation could be improved if teachers receive feedback from group discussion or students performance data. These authors were mainly interested in the teachers' perception of the *difficulty* and not so interested in actual item *heuristic* calibration.

van der Watering and van der Rijt (2006) compare the percentage of correct answers with the estimation of teachers and

students. Their work was carried out at the Faculty of Law of a Dutch University and involved 223 students and 17 teachers. Teacher and students were asked to classify the items in three categories: *easy, not-easy-not-difficult* and *difficult*. They conclude that teachers' estimation of the difficulty of the whole test is appropriate; but that they fail to estimate two thirds of the assessment items and tend to overestimate student performance. They also conclude that students tend to underestimate their performance. These authors were also mainly interested in the teachers and students perception of the *difficulty* and not so interested in actual item *heuristic* calibration.

In a very interesting pilot study Wauters, Desmet, and van Den Noortgate (2012) compare six different estimations of the difficulty: (1) IRT calibration based on the study data (using Rasch 1PL model), (2) proportion correct (CTT), (3) learner feedback, (4) expert rating, (5) one-to-many comparison based on learners' judgment, (6) one-to-many comparison based on experts' judgment, and (7) the Elo rating system. Results indicate that proportion correct has the strongest relation with IRT-based difficulty estimates, followed by student estimation. The experiments were based on a 318 students population (secondary education) and 13 teachers. The topic was Linguistic and Literature. Authors explicitly indicate that no generalization could yet be made to other domains. The results of our study are slightly different but mainly consistent with these findings. In this paper we limit our study to (1) IRT calibration, (2) proportion correct, (3) student rating, (4) teacher rating; which are the most promising techniques according to previous results.

3. The SIETTE assessment system

SIETTE (System of Intelligent Evaluation using Tests) is a web-based system for student knowledge diagnosis (Conejo et al., 2004). This tool is used by teachers as an academic resource either for *formative* or *summative assessment* (Black & William, 2009). The system is currently regularly used at Malaga University, where to is linked to the whole University learning management system. It is also used remotely by different lecturers at the Polytechnic University of Madrid (UPM), the Spanish National University for Distance Education (UNED), Cordoba University (UCO), etc. SIETTE is used in several degree courses such as the B.Sc. and M.Sc. in Computer Science, B.Sc. and M.Sc. in Telecommunications, the M.Sc. in Forestry Engineering among others. Courses such as Programming, Compiler Construction, Databases, Software Engineering, Logic, Statistics, Physics, Botanic, Zoology, Law, English as a second language, etc. frequently use SIETTE. Most of its content is in Spanish for higher education. Its knowledge base contains around 190 courses, 1000 tests, 27,000 items and 30,000 users. Since 2002 when we begun to record, 196,000 test sessions has been taken.

It can be used as an autonomous tool or as a diagnostic module in other web-based environments. The system provides web-services for interacting with other e-learning systems and learning platforms like Moodle.

SIETTE allows the administration of several types of tests. Firstly, conventional tests where the evaluation is done according to heuristics such as the percentage of student success in answering, or the points he/she has obtained. In addition, the system allows the administration and automatic calibration of IRT-based tests, including Computer Adaptive Tests (CAT).

We have implemented several mechanisms to ensure the system security, such as access by username/password, access restrictions according to user groups, IP address, date, location, etc.; to avoid cheating, such as random posing of questions, isomorphic item generation from templates, etc.; and several criteria for item and hints selection, feedback, etc.

The tasks supplied by SIETTE are test items. Three types of items can be distinguished in terms of the response format. These are called internal items. Other tasks format are transformed into these types:

- Multiple-choice items: These items are also composed of a stem and set of choices, equal or greater than two. In this case, students can either leave the item blank or select only one choice.
- Multiple-response items with independent choices: These items have the same format as the former, but in this case, students can select more than one choice (or leave it blank). These items allow partial credit to be awarded to students. This means that the item could be partially correct if a student selects some correct choices.
- Open answer items: In these items, students have to write the answer (or answers) to a given stem. These kinds of items are corrected using patterns. SIETTE manages several types of patterns. The type selected for one item is configured at item construction. Different patterns can model correct and incorrect responses. This is useful in *formative assessment* when feedback is shown after the item correction, since this feedback can be adapted to the specific student response.

SIETTE supports composed items, external items and interactive items where the student must perform some actions in order to solve the tasks. However all these formats are finally transformed into one or more underlying internal items.

From a conceptual point of view the system architecture fits in with the Evidence-Centered Design (ECD) approach to constructing educational assessments in terms of evidentiary arguments (Mislevy, Almond, & Lukas, 2004) the main goal of this proposal is to provide a framework to obtain inferences from what students say or do. In SIETTE the three models can be clearly distinguished:

- The student model: It is an overlay model formed by knowledge probability distribution rep-resenting the student knowledge in the concepts assessed.
- The task model: This model is composed of all the types of items available in SIETTE. It is only in charge of capturing the student response.
- The evidence model: This model uses the student's response supplied by the task model and applies an inference procedure. This procedure determines the concepts whose knowledge can be updated. This can be done in terms of the relationships existing between the concepts. Once the inferences are made, the corresponding knowledge probability distributions of the student model are updated.

From a functional point of view the system architecture contains three elements:

- *The assessment framework*, where students can take tests. Following an authentication process, and the selection of the subject and test to take, the student answers a set of questions. Fig. 1a shows a question posed and incorrectly answered. The upper bar indicates the number of questions previously posed, its correctness (optionally) and a navigation tool that can allow students to go back and forth (optionally). This question requests to identify a tree and shows the correction afterwards (optionally). A green/red mark indicates a selected/unselected correct response, while the red cross indicates the student incorrect response.
- *The editor*using this tool, teachers can create questions and define tests In SIETTE the content is structured into courses. The curriculum of each course is composed of hierarchically structured concepts. Items are linked to the concepts they

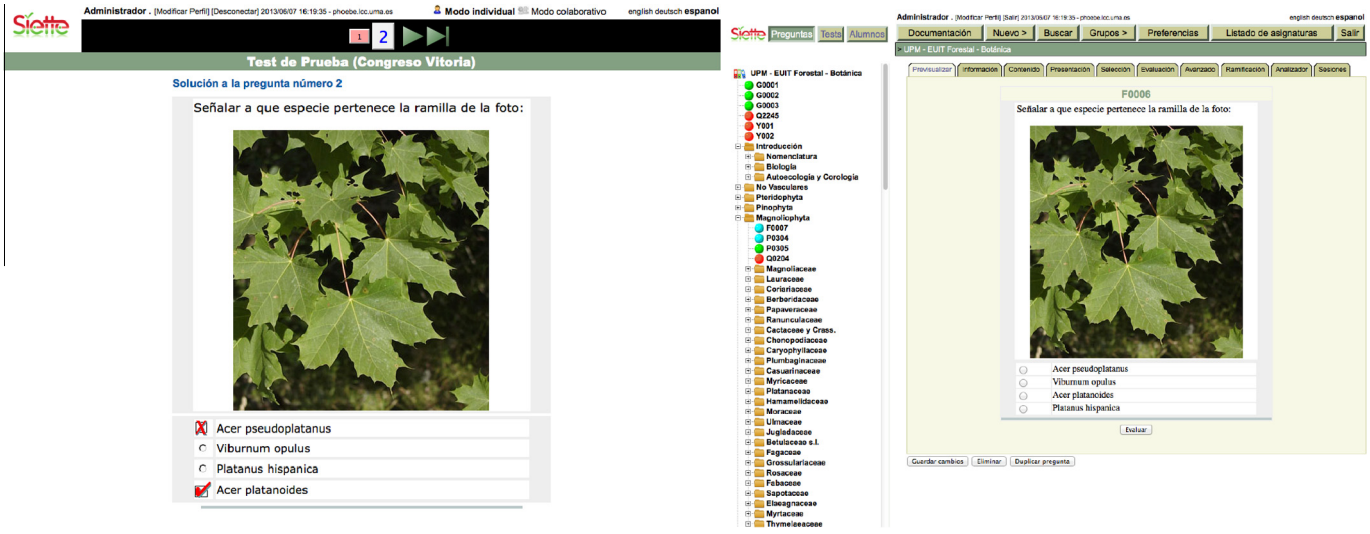

Fig. 1. (a) The student assessment framework. (b) The teacher editor.

assess. Using the editor the teacher can define the stem and answers, and define other parameters like maximum time exposure, presentation and selection information, question metadata, etc. Tests are defined according to different questions selection criteria, evaluation models, item exposure, access constraints, etc. (Fig. 1b).

- *The analyzer*: A set of tools integrated into the editor interface. Using it, teachers can review their students' performance in the tests and obtain summary information about group performance, test characteristics, like descriptive statistics of results, Cronbach alpha, and other indicators. For each item, the teacher can inspect different values like the option selection distribution, (polytomous) item characteristic curves, point biserial correlations, etc. It provides access to the calibration engine, and export data for external processing. Fig. 2 shows some screenshots.

There are many features of SIETTE that cannot be described here. The interested reader is redirected to the system wiki pages. The system is available at: http://www.SIETTE.org.

4. The experiments

As mentioned before, the main goal of these experiments is to explore whether we can trust in human experts to determine the tasks difficulties. We will compare the estimations made by the human experts (in our case, the teachers) with the data inferred by applying test theories. These data-driven *difficulties* will also be compared with students' estimations. Additionally, we will explore the degree of internal coherence among estimations made by a group of teachers and equally for the estimations made by a group of students.

We have conducted three empirical experiments with three different courses and student populations. The same procedure was followed in all three and consisted of the following steps:

1. Instructors of the course constructed a set of test items and a test specification using the SIETTE test editor.
2. Once the items were constructed and reviewed by all of them, each teacher provided an estimation of the difficulty of each item.
3. The test was administrated to the corresponding student population through the SIETTE assessment framework.
4. For each test item, each student had to give us his/her estimation of the item difficulty.
5. After being administrated, all the test items were calibrated according to 1PL, 2PL and 3PL IRT models, and also according to CTT (proportion of correct answers).

4.1. The botany test

4.1.1. Experimental design

The first experiment was conducted with students from the M.Sc. in Botany (Polytechnic University of Madrid, Spain). This test was part of the final qualification of the semester. Students were around 19 years old. They took a test of 99 items, where two types of items could be found: multiple-choice items and multiple response items with independent choices. Items were presented randomly to each student and choices were also randomly ordered to avoid possible cheating. A total of 81 students took the test, which was scored using a point-based criterion. Accordingly, each item was assigned one point if answered correctly. Otherwise, if the answer was incorrect a negative score was assigned. When the item was left blank, no score was awarded.

Once students had finished the test, they were invited to evaluate anonymously the difficulty of each item in a discrete scale between 0 and 10. This scale is the most commonly used in Spain to evaluate students. A total of 13 individuals provided us with their personal estimations. Four *Botany* course teachers provided us with their estimations of the difficulty of each item. We should mention that no one (neither teachers nor students) asked us to clarify the meaning of the concept "item *difficulty*". They were told only that they had to express their estimation on a scale of 0–10. It should be noted that this test was administrated in a controlled environment, i.e. in the laboratories of the school. This was crucial in this case since the test results were taken as part of the student's final qualification.

After test administration, we calibrated the items using students' performance. First we carried out the calibration according to the IRT models, i.e. using 3PL, 2PL and 1PL ICC functions. To this end we used Multilog (Thissen, 2003, which is one of the most popular tools for this purpose. This program uses the Marginal Maximum Likelihood item parameter estimation technique.

We also calibrated the items according to CTT. As mentioned in Section 2.5, difficulty in CTT is defined as the portion of students who answered successfully. Strictly speaking, this definition does

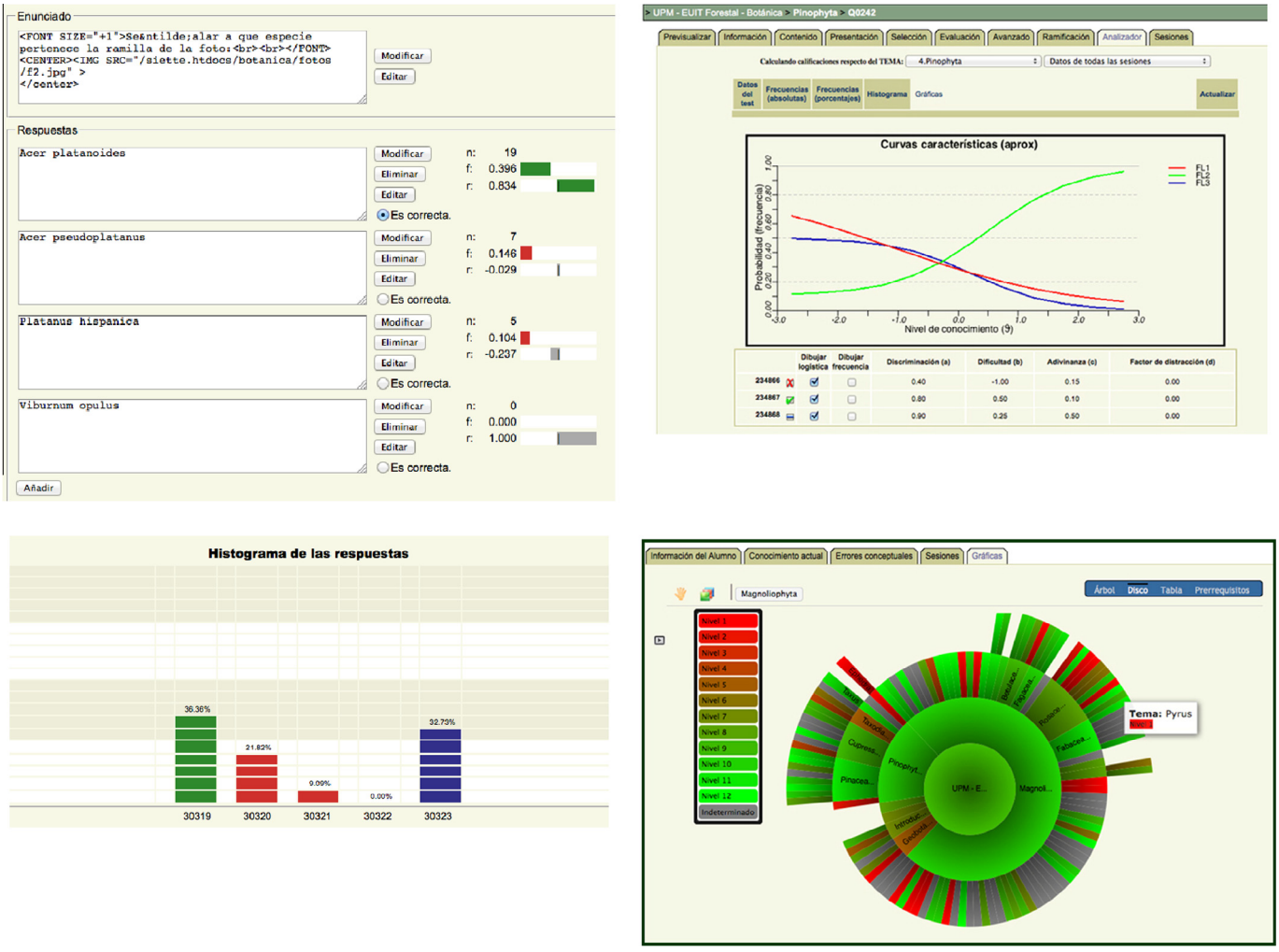

Fig. 2. The SIETTE analyzer.

Table 1
Cronbach's alpha values for the Botany test.

	Cronbach's alpha coefficient
Among teachers	0.73
Among students	0.85
Teachers vs. students	0.78
Among test items	0.96

not correspond to difficulty as much as to the *easiness* of the item. Thus, to obtain the *difficulty* we only have to substrate one from the *easiness* value.

4.1.2. Results

First, we needed to know whether the difficulty estimations obtained from teachers and students were consistent. To this end, we computed Cronbach's alpha coefficient, which is a consistency measurement computed using individuals' estimations. In general, values greater than 0.70 are considered acceptable. Table 1 shows Cronbach values for this experiment. The first row refers to the internal consistency among teachers' difficulty estimations. The second row the reliability of student estimations and the third measures the consistency between the means of these estimations, i.e. the mean of teachers' estimations vs. the mean of students' estimations. As can be seen, even though all results suggest an acceptable level consistence, the students' estimations exhibit a higher (internal) degree of coherence. Ultimately, we observed that the internal consistency within the test items is very high.

Table 2
Results of paired *t*-tests for the Botany test.

	Mean of differences	Confidence interval	*p*-value
Students vs. teachers	−0.085	−0.35, 0.16	0.5275
Students vs. CTT	0.380	0.01, 0.74	0.0393
Students vs. 3PL	−0.463	−0.82, −0.10	0.0113
Students vs. 2PL	0.080	−0.18, 0.34	0.5544
Students vs. 1PL	0.233	−0.06, 0.53	0.1248
Teachers vs. CTT	0.295	−0.13, 0.72	0.1794
Teachers vs. 3PL	−0.548	−0.95, −0.14	0.0085
Teachers vs. 2PL	−0.004	−0.36, 0.35	0.9790
Teachers vs. 1PL	0.148	−0.23, 0.53	0.4428

Using the matrix of students' responses to the test items, we calibrated the items with the three IRT-based parametric models. We also compared estimations and calibration results using a paired *t*-test with a 95% confidence level. This test compares two paired sets to determine whether they differ from each other significantly. The null hypothesis of paired *t*-test is that the mean of the differences is equal to zero. The results of all these tests are shown in Table 2. The second column contains the mean of the differences between the pair estimation-difficulty calibration, the

third column is the confidence interval of this mean; and the last column is the *p*-value of the null hypotheses.

As can be seen, results suggest we cannot reject the hypotheses that students' estimation and difficulties calibrated in the 2PL and 1PL models are similar (p-value > α, with $\alpha = 0.05$). However, the similarity between students' estimations and the other two models cannot be affirmed. We could say the same for the comparison between teachers' estimations and calibrations in the 2PL and 1PL IRT-based models, and even for the calibration results according to CTT. Nonetheless, the null hypothesis can be clearly rejected when comparing teachers and 3PL model. The similarity between estimations made by teachers and students cannot be denied.

We also computed Pearson's correlation coefficient between students' estimations and difficulties calibrated by using 2PL ($r = 0.64$) and 1PL ($r = 0.68$). These results indicate that there is a large positive correlation. We did the same for teachers' estimations and compared them to calibrated values obtained by using the 2PL model ($r = 0.43$), 1PL ($r = 0.49$) and CTT ($r = 0.50$).

Finally, Figs. 3 and 4 represent paired relationships using scatterplots. Fig. 4 depicts the association between students' mean estimations and the difficulty calibration results in the four models: CTT, 3PL, 2PL and 1PL. Fig. 5 illustrates the same comparison, replacing the students' mean for that of the teachers'.

4.2. The LISP test

4.2.1. Experimental design

This experiment was carried out is part of the academic evaluation of an Artificial Intelligence and Knowledge Engineering course, a component of the fourth year of the Computer Science Engineering degree in the University of Málaga (Spain). Students of this course were around 21 years old and had to pass this test

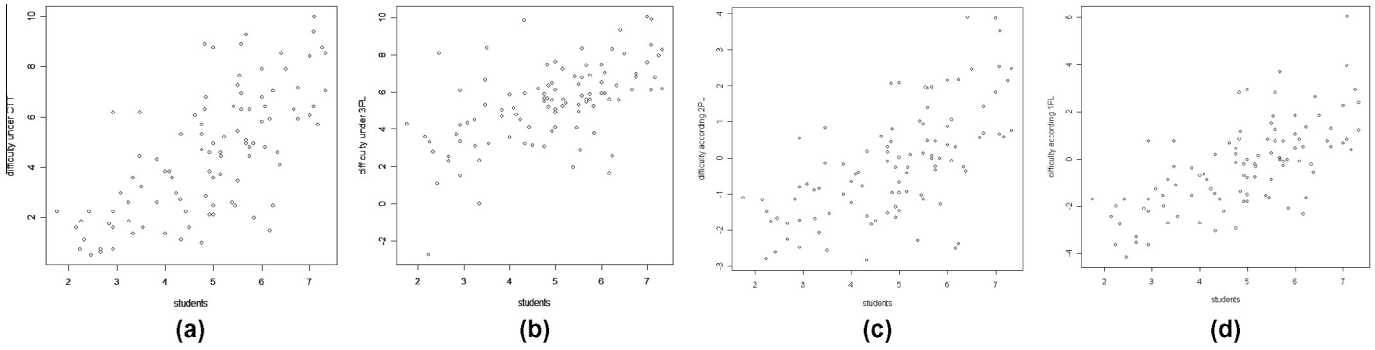

Fig. 3. Scatterplot from the *Botany* test data, comparing students' estimations with calibration results in the models: (a) CTT, (b) 3PL, (c) 2PL, (d) 1PL.

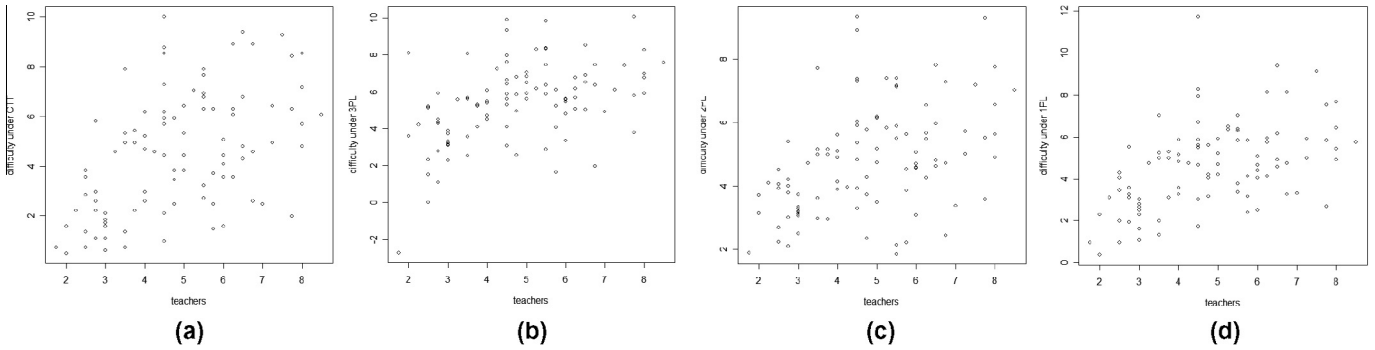

Fig. 4. Scatterplot from the *Botany* test data, comparing teachers' estimations with calibration results in the models: (a) CTT, (b) 3PL, (c) 2PL, (d) 1PL.

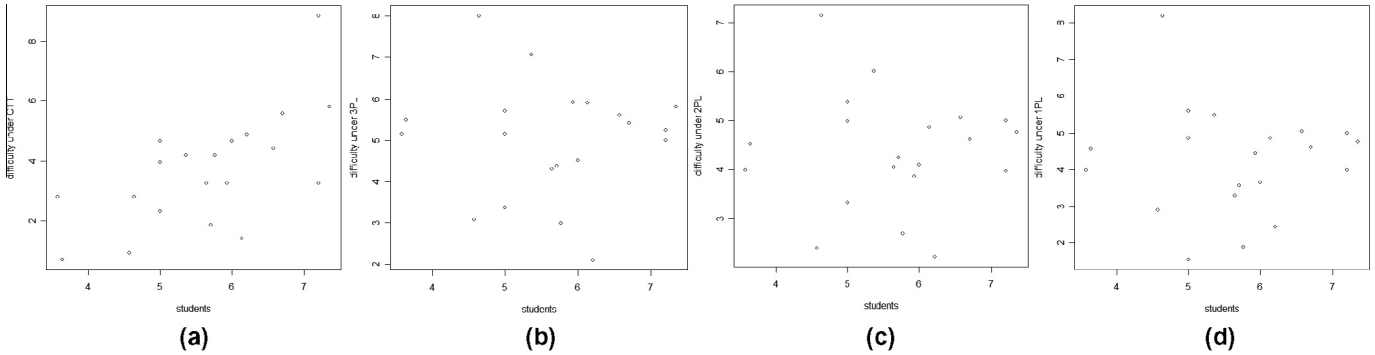

Fig. 5. Scatterplot from the *LISP* test data, comparing students' estimations with calibration results in the models: (a) CTT, (b) 3PL, (c) 2PL, (d) 1PL.

in order to pass the whole course. The test is usually composed of twenty multiple-choice items with only three choices. In this examination session, the students had a time limit of 25 min to answer to all items. The experiment was also carried out in a controlled environment, that is, in the laboratories of the school. The course teachers designed the test in such a way that, at the end of the test, the correction of all items was shown by SIETTE. The test was scored using points, in the same way as the former experiment.

In this case students were also invited to give their estimations about the item difficulties. None of the students asked us to clarify the meaning of difficulty. Once again we only provided the scale in which this value should be expressed, i.e. in the discrete interval [0,10]. Even though 43 students took the test, only 14 individuals gave us their anonymous estimations. We also asked the three course teachers to give us their individual estimations.

4.2.2. Results

As in the former experiment, we analyzed the internal consistency of prior data (Table 3). In this case, consistency among students' difficulty estimations was found to be similar to that of the former student sample and, accordingly, can be considered good enough. However, Cronbach's alpha of the teachers' estimation is under the threshold and therefore suggests lack of internal coherence. For this reason, we have included in this table a separate analysis comparing individual estimations of the three teachers (identified by T1, T2 and T3) with the means of students' difficulty estimations. These results are shown in the fourth, fifth and sixth rows, and suggest that the estimation made by teacher T2 is consistent with student's one. In addition, we checked the coherence of all estimations (teachers and students together) and, in spite of the results for the teacher group, we obtained the best alpha coefficient of the estimations, i.e. 0.85. Finally, the last row of Table 3 shows that the intra-test coherence is good.

After the calibration process, we carried out the paired *t*-tests as in the former experiment (Table 4). Regarding the students' estimation, we can clearly reject the equivalence with difficulty values under CTT and 2PL and 1PL IRT-based models. However, the pair compared with the 3PL calibration results is not statistically significant at the 5% level, and therefore we cannot reject the null hypothesis. After that, we computed the correlation coefficient between the student group estimation and 3PL difficulties ($r = 0.64$). This result suggests a strong and positive correlation.

Due to the lack of coherence among teachers' *difficulty* estimations, we have done the *t*-tests comparing each teacher's estimation with the calibration results. Data suggest that estimations made by teacher T3 are not equivalent to any item calibration model. With respect to teacher T2, although the paired *t*-tests do not reject the null hypothesis regarding the 3PL model, Pearson's coefficient ($r = 0.02$) denotes a very low correlation between both sets of data.

Table 3
Cronbach's alpha values for the *LISP* test.

	Cronbach's alpha coefficient
Among teachers	0.70
Among students	0.89
Teachers vs. students	−0.42
T1 vs. students	−0.20
T2 vs. students	−0.07
T3 vs. students	−0.64
Among test items	0.89
Among teachers	0.70

Table 4
Results of paired *t*-tests for the *LISP* test.

	Mean of differences	Confidence interval	*p*-value
Students vs. CTT	1.975	1.29, 2.66	7.91E−06
Students vs. 3PL	0.651	−0.18, 1.49	0.1215
Students vs. 2PL	1.303	0.54, 2.06	0.0019
Students vs. 1PL	1.432	0.56, 2.30	0.0026
T1 vs. CTT	1.313	0.00, 2.61	0.0489
T1 vs. 3PL	−0.010	−1.03, 1.01	0.9831
T1 vs. 2PL	0.641	−0.29, 1.57	0.1664
T1 vs. 1PL	0.770	−0.20, 1.74	0.1141
T2 vs. CTT	1.813	0.84, 2.78	0.0009
T2 vs. 3PL	0.489	−0.51, 1.49	0.3222
T2 vs. 2PL	1.141	0.25, 2.02	0.0143
T2 vs. 1PL	1.270	0.27, 2.26	0.0148
T3 vs. CTT	2.113	0.87, 3.35	0.0020
T3 vs. 3PL	0.789	0.04, 1.53	0.0397
T3 vs. 2PL	1.441	0.69, 2.18	7E−04
T3 vs. 1PL	1.570	0.81, 2.33	0.0003

t-Test results for T1 teacher do not suggest the null hypothesis rejection regarding IRT-based models. This result is greater for the 3PL model. Pearson coefficients for these comparisons are: 3PL, $r = 0.38$; 2PL, $r = 0.50$; and 1PL, $r = 0.46$. This means that there is a strong positive correlation between T1 estimations and difficulties under 2PL. The others only indicate an acceptable degree of correlation.

Finally, these relationships between students' and individual teacher's mean estimations and the results of calibration have been represented graphically by using scatterplots in Figs. 5 and 6.

4.3. Fundamentals of programming test

4.3.1. Experimental design

This experiment involved teachers and students of a Fundamentals of Programming course corresponding to the second semester of the first year of B.Sc. in Telecommunications in the University of Málaga. The course has around 300 students per academic year (individuals around 18 years old). The course teachers decided to offer their students a new activity in order to prepare them for the final exam. Note that the final exam is composed of a test of 15 multiple-choice items with three choices. This new activity consists of constructing a formative test using SIETTE. The main goal of this test was to provide students with an environment to train for the exam from their homes. We call this type of test open test. In the studies described in Guzmán, Conejo, and Pérez-de-la-Cruz (2007a), empirical evidence suggests that these tests are useful for facilitating the student learning process.

The open test of this experiment had several restrictions:

- Each student was allowed to take the test only once a day (this restrictive facility is provided by SIETTE and is configured during the test elicitation process).
- Once the test was finished, the corrections were not shown and only the final score was supplied to the student. This restriction was included to force the students to try to complete the test, rather than simply copying the correct answers to the items. We have observed in other experiments that many students adopted this strategy. Instead of doing the actual test themselves, they wanted only to see the questions and the correct answers.
- The first time a student took a test, he/she had to submit a difficulty estimation of all the test items. Once again, students did not ask questions about the notion of difficulty. We only provided the scale in which they had to estimate the difficulties, i.e. from 0 to 10. We should point out that we informed the stu-

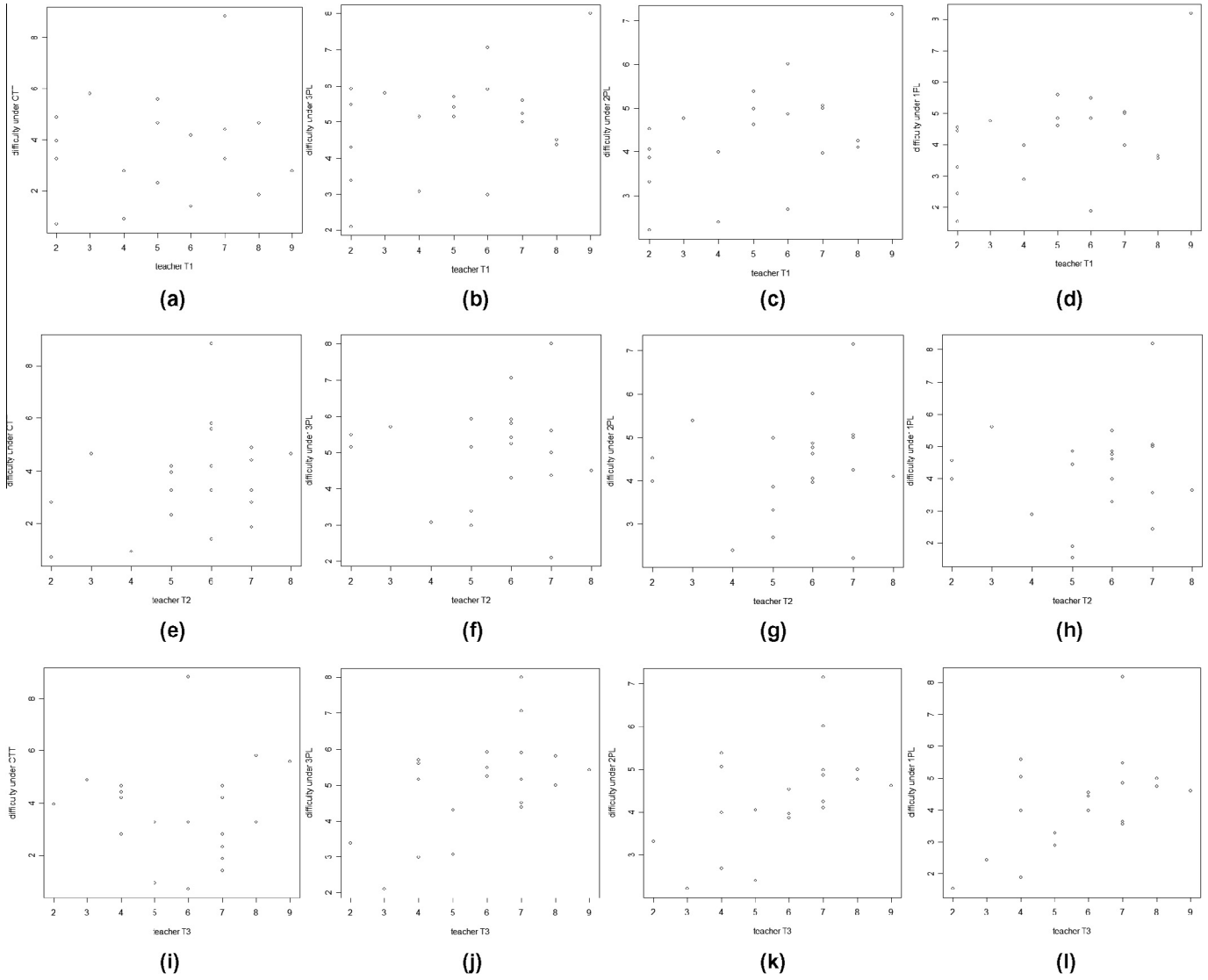

Fig. 6. Scatterplot from the *LISP* test data, comparing teachers' estimations with calibration results in the models: Teacher T1: (a) CTT, (b) 3PL, (c) 2PL, (d) 1PL. Teacher T2: (e) CTT, (f) 3PL, (g) 2PL, (h) 1PL, Teacher T3: (i) CTT, (j) 3PL, (k) 2PL, (l) 1PL.

dents that correction would only be available for those individuals who gave us these estimations.

In order to control access to the test, we gave permission only to those students who requested it from the course teacher. Each student had to supply his/her full name and email address. Once all this information was collected, the teachers gave it to us and we generated the pair username/password necessary to access to the test. When all students were registered, the username, his/her password and the instructions the students had to follow in the test were automatically generated via email. The test consisted of 20 items in which the format and the item type was the same as that in the exam.

Students could take the test once a day for a week. Three days before the exam, the test specification was changed. It contained the same items but this time after each one the correction was shown. Once again, we restricted this test to only those students who had previously given the difficulty estimations.

A total of 233 sessions were collected from 103 individuals who initially participated in this experiment. Only 42 of these individuals gave us their difficulty estimations. The three course teachers also gave us their personal difficulty estimations. As in the former experiments, nobody asked us about the definition of item difficulty. The only information we provided was again the scale in which the values should be expressed: [0,10].

Table 5
Cronbach's alpha values for the *Fundamental of Programming* test.

	Cronbach's alpha coefficient
Among teachers	0.70
Among students	0.89
Teachers vs. students	−0.42
T1 vs. students	−0.20
T2 vs. students	−0.07
T3 vs. students	−0.64
Among test items	0.89

4.3.2. Results

Consistency analysis using Cronbach's alpha coefficient (Table 5) shows that there is an acceptable level of agreement among teachers' difficulty estimations. This value is even better among students' estimations. We also compared the means of both sets of

Table 6
Results of paired *t*-tests for the *Fundamental of Programming* test.

	Mean of differences	Confidence interval	*p*-Value
Students vs. Teachers	−0.202	−1.17, 0.77	0.6690
Students vs. CTT	1.156	0.40, 1.91	0.0046
Students vs. 3PL	0.459	−0.34, 1.25	0.2450
Students vs. 2PL	1.214	0.57, 1.85	0.0008
Students vs. 1PL	1.146	0.39, 1.89	0.0047
Teachers vs. CTT	0.954	−0.01, 1.92	0.0529
Teachers vs. 3PL	0.257	−0.92, 1.43	0.6538
Teachers vs. 2PL	1.012	0.11, 1.91	0.0293
Teachers vs. 1PL	0.944	−0.01, 1.90	0.0535

estimations, i.e. teachers vs. students, but no consistency was found. For this reason, we analyzed teachers' estimations separately. We did not find any correlation between the mean of teacher's and students' estimations. Regarding test item internal consistency, once again, the coefficient was high.

Table 6 contains the paired *t*-tests done after completion of the calibration processes. Even though Cronbach alpha indicates that there is no consistency among the means of teachers' and students' difficulty, the paired *t*-test suggests we cannot reject the hypothesis of similarity between the pairs of difficulty estimation means. Nonetheless, Pearson coefficient denotes a low negative correlation ($r = -0.24$).

With respect to the comparison between the mean students' estimation and the four models, the similarity with CTT, 2PL and 1PL can be clearly rejected. Nevertheless, we cannot reject the null hypothesis for the 3PL model. For this last model, the correlation with student group estimation is positive and acceptable ($r = 0.46$).

When comparing the calibration results with teachers' estimations, the null hypothesis rejection is at the limit for the CTT, 2PL and 1PL models and accordingly, we do not have enough evidence to support the teachers' estimations equivalence with the calibration results of these models. Nonetheless, the evidence does suggest that we cannot reject the similarity with the calibration results in the 3PL model. In spite of that data, the correlation value is positive but very low ($r = 0.07$) and, as a consequence, we cannot affirm the similarity between the difficulties in this model and teacher's estimations.

Finally, as in the two former experiments, Figs. 7 and 8 illustrate the scatterplots representing the associations between students' estimations and teachers' estimation vs. the calibration results.

5. Conclusion

The notion of task *difficulty* is certainly a subjective one. In the literature we can find many systems that use estimations of *difficulty* based on the values given by human experts.

However, our study suggests that human estimations are not realistic. We have performed three experiments. In them we have varied the subject matter evaluated, the student sample and the experts (teachers) who estimated the *difficulty*. We have also tried to isolate the most important (from the student knowledge perspective) component of *difficulty*, i.e. the content *difficulty*. For this purpose we have used simple test items instead of complex tasks. We consider that in complex tasks the influence of other difficulties is higher, and this would have perhaps affected the results of the experiments.

According to the first experiment, students' estimations are better than the those of the teachers' in terms of correlation, when they are compared to *difficulties* inferred from the 2PL and 1PL. models. We could say the same for the second experiment, but this time instead of comparing to the 2PL and to the 1PL, comparing it to the 3PL. Moreover, in this case, the evidence suggests that

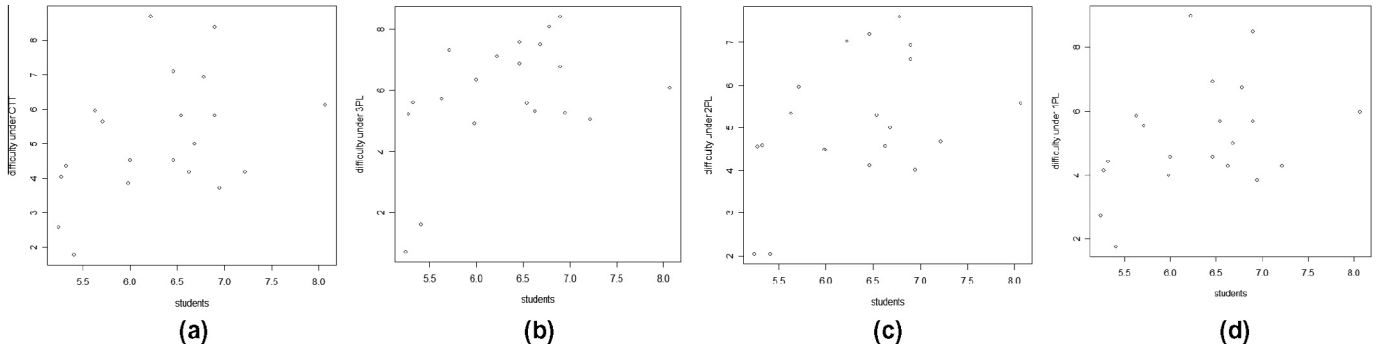

Fig. 7. Scatterplot from the *Fundamental of Programming* test data, comparing students' estimations with calibration results in the models: (a) CTT, (b) 3PL, (c) 2PL, (d) 1PL.

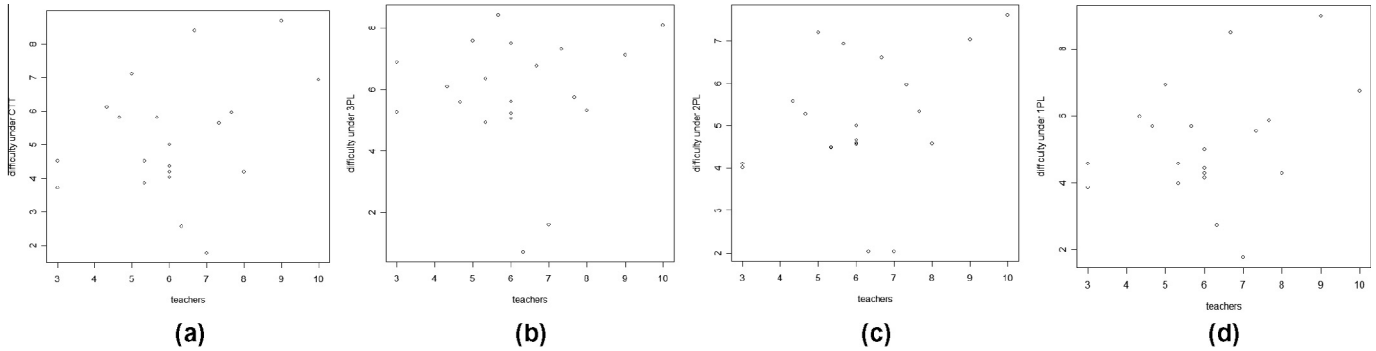

Fig. 8. Scatterplot from the *Fundamental of Programming* test data, comparing teachers' estimations with calibration results in the models: (a) CTT, (b) 3PL, (c) 2PL, (d) 1PL.

teachers' *difficulty* estimations are not valid and are even inconsistent among all the three teachers. Finally, the third experiment supports the results obtained in the former. The students' estimations fit in better with the 3PL-based item difficulties. This last experiment suggests again that teachers' estimations are not accurate.

In summary, the three experiments we have carried out support the hypothesis that human based estimations of difficulty are not consistent with those obtained through data-driven techniques. There is also some evidence that favors the hypothesis that students' estimations are better than teachers' ones. These results are in line with those previously obtained by van der Watering and van der Rijt (2006), that concludes that "*students are better estimators of item difficulty* and that*teachers are able to estimate the difficulty levels correctly for only a small proportion of the assessment items*". Wauters et al. (2012) also concludes that student estimation are slightly better, but as they explained*:* "*It needs to be considered that the estimation by means of learner feedback is based on a larger sample than the estimation by means of expert rating, which could explain the difference between learner feedback accuracy and expert rating accuracy*". This limitation also applies to this study.

Finally, we have verified what other authors have suggested, that is, the high correlation between difficulty values computed using the four psychometric models used in this study. Correlation is especially meaningful when comparing 1PL and CTT, with Pearson's coefficient values higher than 0.90. This is similar to the result obtained by Wauters et al. (2012). In our research we have also explored the relation with other IRT models like 2PL and 3PL. Our results indicate that difficulty parameters in those models differ more from the *difficulty* parameter in CTT, but are closer to actual human estimations.

Given the empirical results described in this work, we consider that it is important to have available well-founded procedures to validate the *difficulty* of the tasks that teachers create for student assessment. Teacher's estimation can be used as an initial estimation to avoid the cold start problem but we need mechanisms to ensure that diagnostic tools measure what they are designed to measure. To this end, assessment models based on IRT provide data-driven techniques for determining task *difficulty*, and have relevant properties such as invariance and reliability. We consider that ITS and AIWBES should benefit from these properties.

References

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

Beck, J., Stern M., & Woolf, B. P. (1997) Using the student model to control problem difficulty. In A. Jameson, C. Paris, C. Tasso (Eds.), *Proceeding of the VI international conference on user modeling* (pp. 277–289).

Black, P., & William, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5–31.

Brusilovsky, P. (2001). Adaptive hypermedia. *User Modeling and User-Adapted Interaction, 11*, 87–110.

Brusilovsky, P., & Peylo, C. (2003). Adaptive and intelligent web-based educational systems. *International Journal of Artificial Intelligence in Education, 13*, 156–169.

Cabada, R. Z., Barrón Estrada, M. L., & Reyes García, C. A. (2011). EDUCA: A web 20 authoring tool for developing adaptive and intelligent tutoring systems using a Kohonen network. *Expert Systems with Applications, 38*(8), 9522–9529.

Chen, C. M., & Duh, L. J. (2008). Personalized web-based tutoring system based on fuzzy item response theory. *Expert Systems with Applications, 34*, 2298–2315.

Chen, C. M., Lee, H. M., & Chen, Y. H. (2005). Personalized E-learning system using item response theory. *Computers and Education, 44*(3), 237–255.

Chrysafiadi, K., & Virvou, M. (2012). Evaluating the integration of fuzzy logic into the student model of a web-based learning environment. *Expert Systems with Applications, 39*(18), 13127–13134.

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. *Journal of Artificial Intelligence in Education, 14*, 29–61.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*, 253–278.

Feng, M., Heffernan, N., & Koedinger, K. (2009). Addressing the assessment challenge with an online system that tutors as it assesses. *User Modeling and User-Adapted Interaction, 19*(3), 243–266.

Geerlings, H., van der Linden, W. J., & Glas, C. A. W. (2013). Optimal test design with rule-based item generation. *Applied Psychological Measurement, 37*(2), 140–161.

Grigoriadou, M., Kornilakis, H., Papanikolaou, K. A., & Magoulas G. D. (2002). Fuzzy inference for student diagnosis in adaptive educational hypermedia. In *Methods and applications of artificial intelligence*, Lecture Notes in Artificial Intelligence, vol. 2308, New York: Springer-Verlag.

Guzmán, E., Conejo, R., & Pérez-de-la-Cruz, J. L. (2007a). Improving student performance using self-assessment tests. *IEEE Intelligent Systems, 22*(4), 46–52.

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

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their application to test development. *Educational Measurement: Issues and Practice, 12*(3), 38–47.

Hambleton, R. K., Swaminathan, J., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage publications.

Hanson, B. A. (2002). IRT Command Language (ICL). Computer software. [Available at <http://www.b-a-h.com/software/irt/icl/index.html>].

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*, 69–81.

Jeremic, Z., Jovanovic, J., & Gaševic, D. (2012). Student modeling and assessment in intelligent tutoring of software patterns. *Expert Systems with Applications, 39*, 210–222.

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. J. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computer and Education, 57*, 1813–1824.

Lee, F. L. (1996) Electronic homework: an intelligent tutoring system in mathematics. Doctoral Dissertation, The Chinese University of Hong Kong Graduate School – Division of Education, November. Available at <http://www.fed.cuhk.edu.hk/en/cuphd/96fllee/content.htm>. Consulted June, 2013.

Mayo, M. Mitrovic, A. (2000) Using probabilistic student model to control problem difficulty. In *Intelligent tutoring systems conference*, Lecture Notes in Computer Science (Vol. 1839, pp. 524–533). Springer-Verlag.

Melis, E., Andres, E., Bündenbender, J., Frischauf, A., Goguadze, G., Libbrecht, P., et al. (2001). Activemath: A generic and adaptive web-based learning environment. *International Journal of Artificial Intelligence in Education, 12*, 385–407.

Millán, E., Descalço, L., Castillo, G., Oliveira, P., & Diogo, S. (2013). Using Bayesian networks to improve knowledge assessment. *Computers and Education, 60*(1), 436–447.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). A brief introduction to evidence-centered design, CSE Report 632, National Center for Research on Evaluation, Standards and Student Testing (CRESST). Center for the Study of Evaluation (CSE), UCLA, Los Angeles, CA.

Muraki, E., & Bock, R. D. (1997). PARSCALE 3: IRT based test scoring and item analysis for graded items and rating scales (computer software). Chicago: Scientific Software International.

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* (Vol. 6787, pp. 243–254). Lecture Notes in Computer Science.

Plake, B. S., & Impara, J. C. (2001). Ability of panellists to estimate item performance for a target group of candidates: An issue in judgemental standard setting. *Educational Assessment, 7*(2), 87–97.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 40*(6), 601–618.

Shute, V. J., Graf, E. A., & Hansen, E. (2005). Designing adaptive, diagnostic math assessments for sighted and visually-disabled students. In L. PytlikZillig, R. Bruning, & M. Bodvarsson (Eds.), *Technology-based education: Bringing researchers and practitioners together*. Greenwich, CT: Information Age Publishing.

Sosnovsky S., Brusilovsky P., Lee, D. H., Zadorozhny, V., & Zhou, X. (2008). Re-assessing the value of adaptive navigation support in E-learning context. In W. Nejdl, J. Kay, P. Pu, & E. Herder. (Eds.), *Adaptive hypermedia and adaptive web-based systems, 5th international conference* (Vol. 5149, pp. 193–203). AH 2008, Springer LNCS.

Thissen, D. (2003). MULTILOG 7: Multiple, Categorical Item Analysis and Test ScoringUsing Item Response Theory. Version 7.0, [Computer software]. Chicago: Scientific Software International Inc., Lincolnwood, IL.

Trella M., Conejo, R., Guzmán, & E Bueno, D. (2003) An educational component-based framework for web-ITS development. In J. M. Cueva et al. (Eds.), *Proceedings of the international conference on web engineering (ICWE 2003)* (Vol. 2722, pp 134-143). Lecture Notes in Computer Science, New York: Springer Verlag.

van der Watering, G., & van der Rijt, J. (2006). Teachers' and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating

the difficulty levels of assessment items. *Educational Research Review, 1*(2), 133–147.

VanLehn, K., & Martin, J. (1997). Evaluation of an assessment system based on Bayesian student modeling. *International Journal of Artificial Intelligence in Education, 8*(2), 179–221.

VanLehn, K. (1988). Student modeling. In M. C. Polson & J. J. Richardson (Eds.), *Foundations of intelligent tutoring systems* (pp. 55–76). Hillsdale, NJ: Lawrence Erlbaum Associates Publishers.

Verdu, E., Verdu, M., Regueras, L., de Castro, J. P., & García, R. (2012). A genetic fuzzy expert system for automatic question classification in a competitive learning environment. *Expert Systems with Applications, 39*(8), 7471–7478.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000) Computerized adaptive testing: A primer. Routledge, 2nd edition.

Wauters, K., Desmet, P., & van Den Noortgate, W. (2012). Item difficulty estimation: An auspicious collaboration between data and judgment. *Computers & Education, 58*(4), 1183–1193.

Weber, G., & Brusilovsky, P. (2001) ELM-ART: An adaptive versatile system for web-based instruction. *International Journal of Artificial Intelligence in Education. Special Issue on Adaptive and Intelligent Web-based Educational Systems*, 12(4), pp. 351–384.

Weeks, J. P. (2010). Plink: An R package for linking mixed-format tests using IRT-based methods. *Journal of Statistical Software, 35*(12), 1–33.

Whorton, S. (2013). Can a computer adaptive assessment system determine, better than traditional methods, whether students know mathematics skills? (Doctoral dissertation, WORCESTER POLYTECHNIC INSTITUTE). Available online at <http://www.wpi.edu/Pubs/ETD/Available/etd-041913-095912/unrestricted/Final.pdf> Retrieved June 2013.

Wilson, D. T., Wood, R., & Gibbons, R. (1991). TESTFACT: Test scoring, item statistics, and item factor analysis [Computer software]. Chicago: Scientific Software International.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items [Computer software]. Chicago: Scientific Software International.