A Web-Based E-Testing System Supporting Test Quality Improvement

Gennaro Costagliola, Filomena Ferrucci, Vittorio Fuccella
Dipartimento di Matematica e Informatica – Università degli Studi di Salerno, Fisciano(SA),Italy
{gcostagliola,fferrucci,vfuccella}@unisa.it

ABSTRACT

In e-testing it is important to administer tests composed of good quality question items. By the term “quality” we intend the potential of an item in effectively discriminating between strong and weak students and in obtaining tutor’s desired difficulty level. Since preparing items is a difficult and time-consuming task, good items can be re-used for future tests. Among items with lower performances, instead, some should be discarded, while some can be modified and then re-used. This paper presents a Web-based e-testing system which detects defective question items and, when possible, provides the tutors with advice to improve their quality. The system detects defective items by firing rules. Rules are evaluated by a fuzzy logic inference engine. The proposed system has been used in a course at the University of Salerno.

Keywords

e-Testing, Computer Aided Assessment, CAA, item, item quality, questions, eWorkbook, Item Response Theory, IRT, Test Analysis, online testing, difficulty, discrimination, multiple choice, distractor.

1. INTRODUCTION

E-testing systems are more and more widely adopted in academic environments combined with other assessment means. Through these systems, tests composed of several question types can be presented to the students in order to assess their knowledge. *Multiple Choice* question type is extremely popular, since, among other advantages, a large number of its outcomes can be easily corrected automatically. The experience gained by educators and the results obtained from several experiments [22] provide some guidelines for writing good multiple choice questions (*items*, in the sequel), such as: “use the right language”, “avoid a big number of unlikely distractors for an item”, etc.

It is possible to evaluate the effectiveness of the items, through the use of several statistical models, such as *Item Analysis* (*IA*) and *Item Response theory* (*IRT*) [8]. They are both based on the interpretation of statistical indicators calculated on test outcomes. The most important of them are the *difficulty* indicator, which measures the difficulty of the items, and the *discrimination* indicator, which represents the information of how well an item discriminates between strong and weak students. More statistical indicators are related to the *distractors* (wrong options) of an item.

An item with a high value for discrimination is a good item, that is, an item that is answered correctly by strong students and incorrectly by weak ones, on average. Furthermore, in this study we regard as more efficient those items whose calculated difficulty tends to be closer to the difficulty guessed by the tutor. In a test, in order to better assess a heterogeneous class with different levels of knowledge, it is important to balance the difficulty of the items: tests should be composed of given percentages of difficult (25%), medium (50%) and easy (25%) items. If the tutor succeeds in giving the desired difficulty level to an item, he/she can more easily construct balanced tests which assess students on the desired knowledge.

Despite the availability of guidelines for writing good items and statistical models to analyze their quality, only a few tutors are aware of the guidelines and even fewer are used with statistics. The result is that the quality of the tests used for exams or admissions is sometimes poor and in some cases could be improved. Although it is almost impossible to compel the tutors to read manuals for writing good test assessment, it is possible to give them feedback on their items’ quality, allowing them to discard defective items or to modify them in order to improve their quality for next use and, at the same time, to learn how to write good items from experience.

This paper presents a Web-based e-testing system which helps the tutors to obtain good quality assessment items. By item quality we intend the potential of an item in effectively discriminating between strong and weak students and in obtaining a tutor’s desired difficulty level. After a test session, the system marks the items: good items are marked with a green light. For poor quality items there are two different levels of alarm: *severe* (red light), for items which should be discarded, and *warning* (yellow light) for items whose quality could be improved. For the latter ones, the system provides the tutor with suggestions for improving item quality. Aware of defective items, and helped by the suggestions of the system, the tutor can discard or modify poor items. Improvable items can be re-used for future tests.

Quality level and eventual suggestions are decided through a rule-based classification [23]. Fuzzy logic has been used in order to obtain a degree of fulfillment of each rule. Rules have been preferred over other frequently used classification methods, such as hierarchical methods [1], K-means methods [15] and correlation methods due to the following reasons:

- *Knowledge availability*. Most of the knowledge is already available, as witnessed by the presence of numerous theories and manuals on psychometrics.
- *Lack of data*. Other types of classification based on data would require the availability of large data sets. Once they have gathered, in such a way to have statistically significant classes to perform data analysis, such methods might be exploited.

The system has been carried out by adding the formerly described features to an existing Web-based e-testing system: *eWorkbook* [5], developed at *University of Salerno*, which has been equipped with an *Item Quality Module*. A first experiment has been carried out in a course at the University of Salerno.

The paper is organized as follows: section 2 gives some concepts about the knowledge on which the system is based. In section 3, the system is defined, following the steps of a classical methodology for fuzzy systems definition. In section 4, we briefly discuss the implementation of the quality module and its integration in the existing e-testing system. Finally, section 5 presents an experiment and a discussion of its results. The paper concludes with a brief survey on work related to ours, several final remarks and a discussion on future work.

2. THE KNOWLEDGE-BASE

Our system makes use of *multiple choice* items for the assessment of students' knowledge. Those items are composed of a *stem* and a list of *options*. The stem is the text that states the question. The only correct answer is called the *key*, whilst the incorrect answers are called *distractors* [22].

Test results can be statistically analyzed to check item quality. As mentioned in the previous section, two main statistical models are available: *IA* and *IRT*. Several studies, such as the one in [20], make a comparison between the two models, often concluding that they can both be effective in evaluating the quality of the items. For our study, IA has been preferred to IRT for the following main reasons: it needs a smaller sample size for obtaining statistically significant indicators and it is easier to use IA indicators to compose rule conditions. The following statistical indicators are calculated by our system for each item answered by a significant number of students:

- *difficulty*: a real number between 0 and 1 which expresses a measure of the difficulty of the item, intended as the proportion of learners who get the item correct.
- *discrimination*: a real number between -1 and 1 which expresses a measure of how well the item discriminates between good and bad learners. Discrimination is calculated as the *point biserial* correlation coefficient between the score obtained on the item and the total score obtained on the test.
- *frequency(i)*: a real number between 0 and 1 which expresses the frequency of the i-th option of the item. Its value is calculated as the percentage of learners who choose the i-th option.
- *discrimination(i)*: a real number between -1 and 1 which expresses the discrimination of the i-th option. Its value is the point biserial correlation coefficient between the result obtained by the learner on the whole test and a dichotomous variable that says whether the i-th option was chosen (yes=1, no=0) by the learner or not.
- a*bstained_freq*: a real number between 0 and 1 which expresses the frequency of the abstention (no answers given) on the item. Its value is calculated as the percentage of learners who didn't give any answer to the item, where allowed.
- a*bstained_discr*: a real number between -1 and 1 which expresses the discrimination of the abstention on the item. Its value is the point biserial correlation coefficient between the result obtained by the learner on the whole test and a dichotomous variable that says whether the learner refrained or not (yes=1, no=0) on the item.

Discrimination and difficulty are the most important indicators. They can be used for both determining item quality and choosing advice for tutors. As experts suggest [16], a good value for discrimination is about 0.5. A positive value lower than 0.2 indicates an item which does not discriminate well. This can be due to several reasons, including: the question does not assess learners on the desired knowledge; the stem or the options are badly/ambiguously expressed; etc. It is usually difficult to understand what is wrong with these items and more difficult to provide a suggestion to improve them, so, if the tutor cannot understand the problem her(him)self, the suggestion is to discard the item. A negative value for discrimination, especially if joined with a positive value for the discrimination of a distractor, is a sign of a possible mistake in choosing the key (a data entry error occurred). In this case it is easy to recover the item by changing the key.

If the difficulty level is too high (>0.85) or too low (<0.15), there is the risk of not correctly evaluating on the desired knowledge. This is particularly true when such values for the difficulty are sought together with medium-low values for discrimination. Furthermore, our system allows the tutor to define the foreseen difficulty for an item. Thus, the closer a tutor’s estimation of item difficulty is to the actual calculated difficulty for that item, the more reliable that item is considered to be. When difficulty is too high or underestimated, this can be due to the presence of a distractor (noticed for its high frequency) which is too plausible (it tends to mislead a lot of students, even strong ones). Removing or substituting that distractor can help in obtaining a better item. Sometimes, the item has its intrinsic difficulty and it can be difficult to adjust it, so the suggestion can be to modify the tutor’s estimation.

As for distractors, they can contribute to a good item when they are selected by a significant number of students. When the frequency of the distractor is too high, there could be an ambiguity in the formulation of the stem or of the distractor. A good indicator of distractors’ quality is their discrimination, which should be negative, denoting that the distractor was selected by weak students. In conclusion, a good distractor is the one which is selected by a small but significant number of weak students.

High abstention is always a symptom of high difficulty for the item. When it is accompanied by a high (not negative or next to 0) value for its discrimination and a low value for item discrimination, it can tell that the question has a bad quality and it is difficult to improve it.

3. THE FUZZY SYSTEM

The system for the evaluation of item quality is *rule-based*: the rules use, as *linguistic variables*, statistical indicators calculated after a test *session*. By this term we mean the time necessary to administer several items to a statistically significant number of students. The value of this number is set in the configuration of the system.

The system works by performing a *classification* of the items. Several classes of items have been identified, and each class is associated to a *production rule*. The *degree of fulfillment* of a rule tells the membership of the item to the corresponding class. The classification is performed by selecting the class for which the degree of fulfilment is the highest.

3.1 Variables and Fuzzyfication

The set of variables used are reported, together with an explanation of their meaning and the set of possible values they can assume (*terms*), in table 1. These variables are directly chosen from the statistical indicators presented in section 2 or derived from them.

Table 1. Variables and Terms

Variable	Explanation	Terms
discrimination	Item’s discrimination (see sec. 2)	Negative, low, high
difficulty	Item’s difficulty (see sec. 2)	Very_low, medium, very_high
difficulty_gap	The difference between the tutor’s estimation of item’s difficulty and the difficulty calculated by the system	Underestimated, correct, overestimated
max_distr_discr	The maximum discrimination for the distractors of an item	Negative, positive
max_distr_freq	The maximum (relative) frequency for the distractors of an item.	Low, high
min_distr_freq	The minimum (relative) frequency for the distractors of an item	Low, high
distr_freq	The (relative) frequency of the distractor with maximum discrimination for an item	Low, high
abst_frequency	The frequency of the abstentions for an item	Low, high
abst_discrimination	The discrimination of the abstentions for an item	Negative, positive

The variables *discrimination* and *difficulty* are the same indicators for item discrimination and difficulty defined in section 2. The same discourse is valid for the variables related to the abstention, *abst_frequency* and *abst_discrimination*. *difficulty_gap* is a variable representing the error in tutor’s estimation of item difficulty. Through the system interface, the tutor can assign one on three difficulty level to an item (easy = 0.3; medium = 0.5; difficult = 0.7). *difficulty_gap* is calculated as the difference between the tutor estimation and the actual difficulty calculated by the system.

Three variables representing the frequency of the distractors for an item have been considered: *max_distr_freq*, *min_distr_freq*, *distr_freq*. Their value is not an absolute frequency, but relative to the frequency of the other distractors: it is obtained by dividing the absolute frequency by the mean frequency of the distractors of the item. In the case of items with five options, as our system has been tested, their value is a real number varying from 0 to 4.

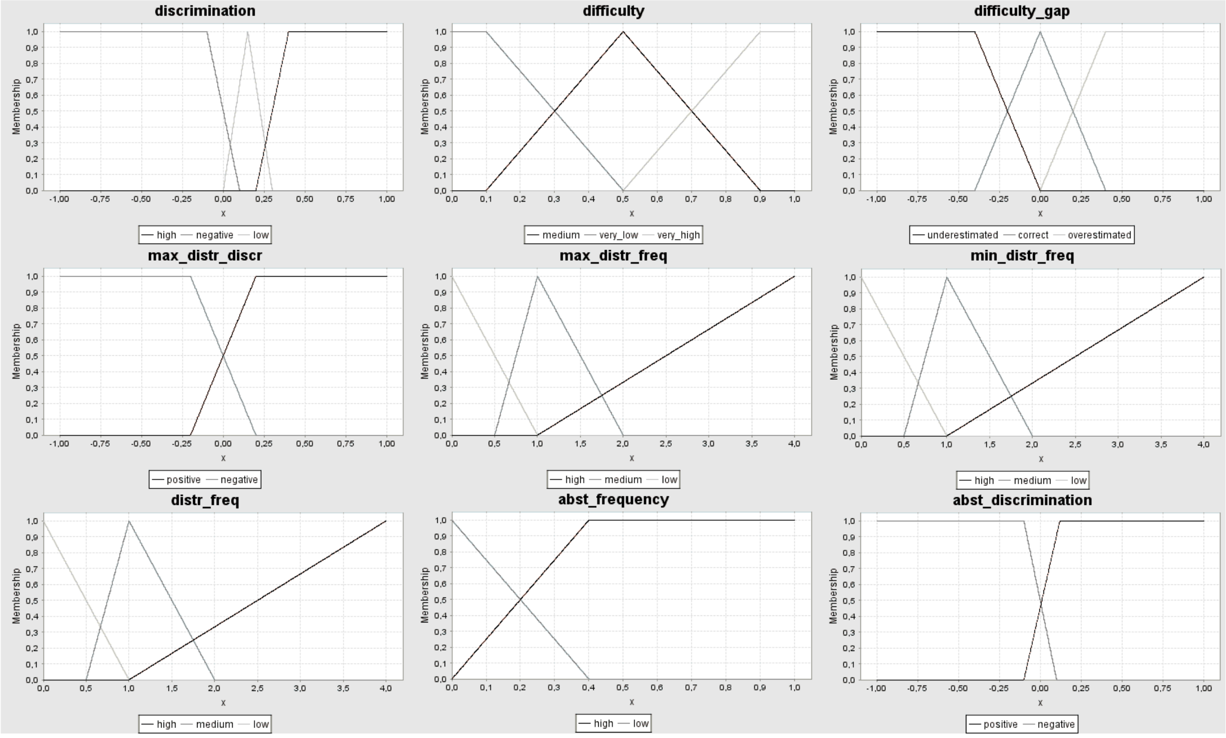


Figure 1. Membership Functions of the Fuzzy Sets

3.2 Membership Functions

As for the membership functions of fuzzy sets associated to each term, *triangular* and *trapezoidal shapes* have been used. Most of the values for the bases and the peaks have been established using the expertise. Only for some variables, the membership functions have been defined on an experimental basis.

While we already had clear ideas on how to define some membership functions, we did not have enough information from the knowledge-base on how to model membership functions for the variables related to abstention (*abst_frequency* and *abst_discrimination*). A calibration phase was required in order to refine the values for the bases and peaks of their membership functions. As a calibration set, test results from the *Science Faculty Admission Test* of the last year (2006) were used. The calibration set was composed of 64 items with 5 options each. For each item, about one thousand records (students answers) were available, even if only a random sample of seventy of them was considered. Test items and their results were inspected by a human expert who identified items which should have been discarded due to low discrimination and anomalous values for the variables related to abstention. We have found 5 items satisfying the conditions above: the mean values for *abst_discrimination* and *abst_frequency* were, respectively, 0.12 and 0.39.

Due to the limited size of the calibration set, the simple method of choosing the peaks of the functions at the mean value, as shown in [1], has been used. When more data will be available, a more sophisticated method will be used for the definition of membership functions, such as the one proposed in [4]. Charts for the membership functions are shown in figure 1.

3.3 Rules

From the verbal description of the knowledge presented in section 2, the rules summarized in table 2 have been inferred. The first three columns in the table contain, respectively, the class of the item, the rule used for classification and the item state. For items whose state is yellow, the fourth column contains the problem affecting the item and the suggestion to improve its quality.

Conditions in the rules are connected using AND and OR logic operators. The commonly-used *min-max* inference method has been used to establish the degree of fulfillment of the rules. All the rules were given the same weight, except for the first one. By modifying the weight of the first rule, we can tune the sensitivity of the system: the lower this value, the higher the probability that anomalies will be detected in the items. Some rules suggest to perform an operation on a distractor. The distractor to modify or eliminate (in case of rules 4, 7 and 10) or to select as correct answer (rule 9) is signaled by

Table 2. Rules for Item Classification

Class	Rule	State	Problem and Suggestion
1	`discrimination IS high AND abst_discrimination IS negative WITH 0.9`	Green	/
2	`discrimination IS low AND abst_frequency IS high AND abst_discrimination IS positive`	Red	/
3	`difficulty IS very_low AND discrimination IS low`	Red	/
4	`difficulty IS very_high AND discrimination IS low AND max_distr_freq IS high`	Yellow	Item too difficult due to a too plausible distractor, delete or substitute distractor *x*.
5	`difficulty_gap IS overestimated AND discrimination IS low`	Yellow	Item difficulty overestimated, avoid too plausible distractors and too obvious answers.
6	`difficulty_gap IS overestimated AND discrimination IS NOT low`	Yellow	Item difficulty overestimated, modify the estimated difficulty.
7	`difficulty_gap IS underestimated AND max_distr_freq IS high`	Yellow	Item difficulty underestimated due to a too plausible distractor, delete or substitute distractor *x*.
8	`difficulty_gap IS underestimated AND max_distr_freq IS NOT high`	Yellow	Item difficulty underestimated, modify the estimated difficulty.
9	`max_distr_discr IS positive AND discrimination IS negative`	Yellow	Wrong key (data entry error), select option *x* as the correct answer.
10	`discrimination IS high AND max_distr_discr IS positive AND distr_freq IS NOT low`	Yellow	Too plausible distractor, delete or substitute distractor *x*.

the system. An output variable *x* has been added to the system to keep the identifier of the distractor.

4. DEVELOPMENT OF THE ITEM QUALITY MODULE AND INTEGRATION IN THE EWORKBOOK SYSTEM

A software module for the evaluation of item quality has been implemented as a Java *Object Oriented* framework. In this way, it would have been easily integrated in any e-testing java-based system. For each item, the module performs the classification, by implementing the following functionalities:

- Implementation of an *Application Programming Interface* (*API*) for the construction of a data matrix containing all the students' responses to the item.
- Calculation of the statistical indicators, as described in section 2.
- Substitution of the variables, evaluation of the rules and choice of the class which the item belongs to.

Implementation of a suitable *API* for obtaining the *state* of an item (*green*, *yellow*, *red*) and, in case of yellow, of the *suggestions* for improving the item. It is worth noting that suggestions can be internationalized, that is, they can easily be translated into any language by editing a text file.

A free java library implementing a complete Fuzzy inference system, named *jFuzzyLogic* [10] has been used. The system variables, fuzzyfication, inference methods and the rules have been defined using *Fuzzy Control Language* (*FCL*) [7], supported by the *jFuzzyLogic* library. The advantage of this approach, compared to a hard-coded solution, is that membership functions and rules can be changed only by editing a configuration file, thus avoiding to build the system again. Data can be imported from various sources and exported to several formats, such as spreadsheets or relational databases. The data matrix and the results can be saved in persistent tables, in order to avoid to perform calculations every time they must be visualized.

4.1 eWorkbook

eWorkbook is a Web-based e-testing system that can be used for evaluating learner's knowledge by creating (the tutor) and taking (the learner) on-line tests based on multiple choice question types. The questions are kept in a *hierarchical* repository. The tests are composed of one or more sections. There are two kinds of sections: *static* and *dynamic*. The difference between them is in the way they allow question selection: for a static section, the questions are chosen by the tutor. For a dynamic section, some selection parameters must be specified, such as the difficulty, leaving the system to choose the questions randomly whenever a learner takes a test. In this way, it is possible with eWorkbook to make a test with banks of items of different difficulties, thus balancing test difficulty, in order to better assess a heterogeneous set of students. eWorkbook adopts the classical *three-tier* architecture of the most common *J2EE* Web-applications. The *Jakarta Struts framework* has been used to support the *Model 2* design paradigm, a variation of the classic *Model View Controller* (*MVC*) approach. In our design choice, Struts works with *JSP*, for the View, while it interacts with *Hibernate* [9], a powerful framework for *object/relational persistence* and query service for Java, for the Model. The application is fully accessible with a Web Browser. No browser plug-in installations are needed, since its pages are composed of standard HTML and *ECMAScript* [6] code.

4.2 Integration

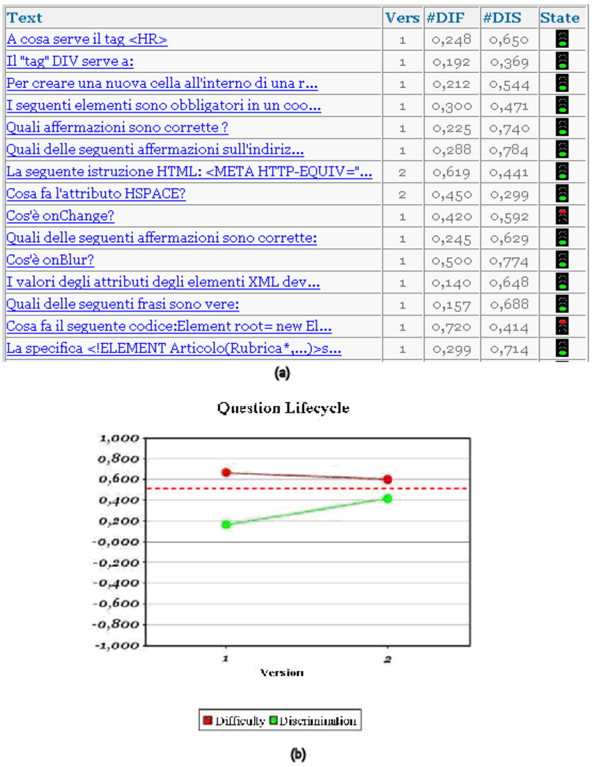
The integration of the new functionalities in eWorkbook has required the development of a new module, named *Item Quality Module*, responsible for instantiating the framework and providing import, export and visualization functionalities. Import of data was performed by reading data from eWorkbook's database and by calling the *API* to fill the data matrix of the framework. The interface for browsing the item repository in eWorkbook has been updated in order to show item's performances (difficulty and discrimination) and state (green, yellow or red). In this way, defective items are immediately visible to the tutor, who can undertake the opportune actions (delete or modify). A screenshot of the item report is shown in figure 2a.

Furthermore, the system has been given a *versioning functionality*: once an item is modified, a newer version of it is generated. Through this functionality, the tutor can analyze the entire lifecycle of an item. In this way, the tutor can have feedback on the trend of statistical indicators over time, making sure that the changes he/she made to the items positively affected their quality. Figure 2b shows the chart of an item improved across two sessions of tests. The improvement is visible both from the increase in the item discrimination (the green line), and in the convergence of the calculated difficulty with the tutor's estimation of the difficulty (the continuous and dashed red lines, respectively).

5. USE CASE IN A UNIVERSITY COURSE

A first experiment has consisted of using the system across two test sessions in a university course, and measuring the overall improvement of the items in terms of discrimination capacity and matching to a tutor's desired difficulty. A database of 50 items was arranged for the experiment. In the first session, an on-line test, containing a set of 25 randomly chosen items, was administered to 60 students. After, items were inspected through the system interface in order to check those to substitute or modify. Once the substitutions and modifications were performed, the modified test was administered to 60 other students.

Figure 3a shows a table, exported in a spreadsheet, containing a report of the items presented in the first test session and their performances. The item to eliminate are highlighted in red, while those to modify are highlighted in yellow. According to the system analysis, 5 out of 25 items must be discarded, while 4 of them must be modified.



Text	Vers	#DIF	#DIS	State
A cosa serve il tag <HR>	1	0,248	0,650	🟢
Il "tag" DIV serve a:	1	0,192	0,369	🟢
Per creare una nuova cella all'interno di una r...	1	0,212	0,544	🟢
I seguenti elementi sono obbligatori in un coo...	1	0,300	0,471	🟢
Quali affermazioni sono corrette ?	1	0,225	0,740	🟢
Quali delle seguenti affermazioni sull'indiriz...	1	0,288	0,784	🟢
La seguente istruzione HTML: <META HTTP-EQUIV="...	2	0,619	0,441	🟢
Cosa fa l'attributo HSPACE?	2	0,450	0,299	🟢
Cos'è onChange?	1	0,420	0,592	🔴
Quali delle seguenti affermazioni sono corrette:	1	0,245	0,629	🟢
Cos'è onBlur?	1	0,500	0,774	🟢
I valori degli attributi degli elementi XML dev...	1	0,140	0,648	🟢
Quali delle seguenti frasi sono vere:	1	0,157	0,688	🟢
Cosa fa il seguente codice:Element root= new El...	1	0,720	0,414	🔴
La specifica <!ELEMENT Articolo(Rubrica*,...)>s...	1	0,299	0,714	🟢

(a)

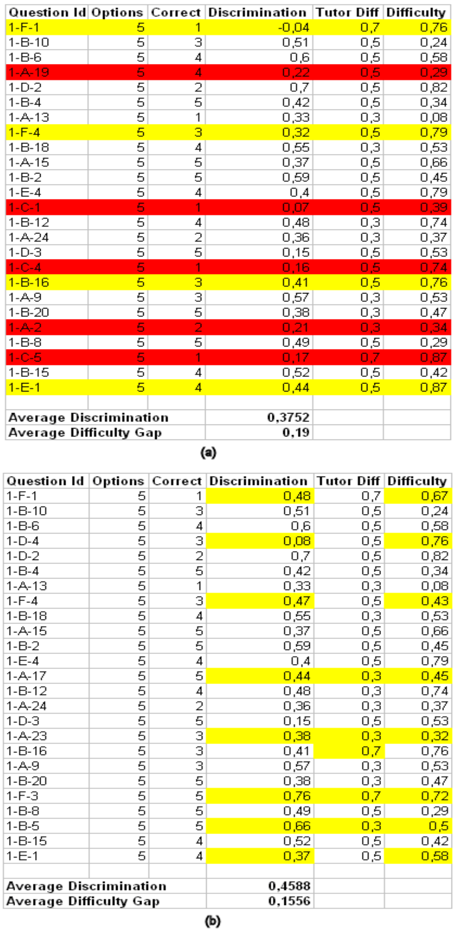
(b)

Figure 2. Screenshots From the eWorkbook System Interface

Actually, among the items to modify, for two of them (those with id 1-F-4 and 1-E-1) the difficulty was underestimated due to a distractor that was too plausible (class 7), which was substituted with a new distractor. In another case (1-B-16), the difficulty was different from that estimated by the tutor, due to the intrinsic difficulty of the item (class 8). The action undertaken was to adjust tutor's estimation of the difficulty.

Lastly, the item with id 1-F-1, with a negative discrimination, presented a suspect error in the choice of the key (class 9). By inspecting the item, the tutor verified that the chosen key was not correct, even though the distractor labeled correct by the system was not the right answer: simply, the item did not have any correct answer. The text of the key was modified to provide the right answer to the stem.

A new test was prepared, containing the same items of the previous, except for the 5 discarded ones, substituted by 5 unused items, and for the 4 modified ones, which were substituted by a newer version of themselves. A new set of sixty students participated in this test. In the analysis of test outcomes, our attention was more focused on the eventual improvement obtained than on the discovery of new defective items.



Question Id	Options	Correct	Discrimination	Tutor Diff	Difficulty
1-F-1	5	1	-0,04	0,7	0,76
1-B-10	5	3	0,51	0,5	0,24
1-B-6	5	4	0,6	0,5	0,58
1-A-19	5	4	0,22	0,5	0,29
1-D-2	5	2	0,7	0,5	0,82
1-B-4	5	5	0,42	0,5	0,34
1-A-13	5	1	0,33	0,3	0,08
1-F-4	5	3	0,32	0,5	0,79
1-B-18	5	4	0,55	0,3	0,53
1-A-15	5	5	0,37	0,5	0,66
1-B-2	5	5	0,59	0,5	0,45
1-E-4	5	4	0,4	0,5	0,79
1-C-1	5	1	0,07	0,5	0,39
1-B-12	5	4	0,48	0,3	0,74
1-A-24	5	2	0,36	0,3	0,37
1-D-3	5	5	0,15	0,5	0,53
1-C-4	5	1	0,16	0,5	0,74
1-B-16	5	3	0,41	0,5	0,76
1-A-9	5	3	0,57	0,3	0,53
1-B-20	5	5	0,38	0,3	0,47
1-A-2	5	2	0,21	0,3	0,34
1-B-8	5	5	0,49	0,5	0,29
1-C-5	5	1	0,17	0,7	0,87
1-B-15	5	4	0,52	0,5	0,42
1-E-1	5	4	0,44	0,5	0,87
Average Discrimination			**0,3752**		
Average Difficulty Gap			**0,19**		

(a)

Question Id	Options	Correct	Discrimination	Tutor Diff	Difficulty
1-F-1	5	1	0,48	0,7	0,67
1-B-10	5	3	0,51	0,5	0,24
1-B-6	5	4	0,6	0,5	0,58
1-D-4	5	3	0,08	0,5	0,76
1-D-2	5	2	0,7	0,5	0,82
1-B-4	5	5	0,42	0,5	0,34
1-A-13	5	1	0,33	0,3	0,08
1-F-4	5	3	0,47	0,5	0,43
1-B-18	5	4	0,55	0,3	0,53
1-A-15	5	5	0,37	0,5	0,66
1-B-2	5	5	0,59	0,5	0,45
1-E-4	5	4	0,4	0,5	0,79
1-A-17	5	5	0,44	0,3	0,45
1-B-12	5	4	0,48	0,3	0,74
1-A-24	5	2	0,36	0,3	0,37
1-D-3	5	5	0,15	0,5	0,53
1-A-23	5	3	0,38	0,3	0,32
1-B-16	5	3	0,41	0,7	0,76
1-A-9	5	3	0,57	0,3	0,53
1-B-20	5	5	0,38	0,3	0,47
1-F-3	5	5	0,76	0,7	0,72
1-B-8	5	5	0,49	0,5	0,29
1-B-5	5	5	0,66	0,3	0,5
1-B-15	5	4	0,52	0,5	0,42
1-E-1	5	4	0,37	0,5	0,58
Average Discrimination			**0,4588**		
Average Difficulty Gap			**0,1556**		

(b)

Figure 3. Report of the Test Sessions

Figure 3b shows the report of the second test session. The values of discrimination and difficulty, changed in respect to the same rows of the session 1 table, are highlighted in yellow.

To measure the overall improvement of the new test, in respect to the previous one, the following parameters were calculated for each of the two tests:

- the mean of the discriminations for the items;
- the mean of the differences |tutor_difficulty – difficulty| for the items of the tests;

As for parameter 1, we have observed an improvement from a value of 0,375, obtained in the first session, to a value of 0,459, obtained in the second session. The percentage of increment is 22,4%. As for parameter 2, we had a decrement in the mean difference between the difficulty estimated by the tutor and the one calculated by the system of 17,8%, passing from a value of 0,19 to 0,156 across the two sessions.

6. RELATED WORK

Several different assessment tools and applications to support blended learning have been analyzed, starting from the most common Web-based e-learning platforms, such as Moodle [17], Blackboard [2], and Questionmark [18]. These systems generate and show item statistics parameters but they do not interpret them, so they do not advise or help the tutor in improving items erasing anomalies revealed by statistics. A model for presenting test statistics, analysis, and to collect students' learning behaviors for generating analysis result and feedback to tutors is described in [12]. IRT has been applied in some systems [11] and experiments [3, 21] to select the most appropriate items for examinees based on individual ability. In [3], the fuzzy theory is combined with the original IRT to model uncertainly learning response. The result of this combination is called *Fuzzy Item Response Theory*.

A work closely related to ours is presented in [13]. It proposes an e-testing system, where rules can detect defective items, which are signaled using traffic lights. It proposes an analysis model based on IA. Statistics are calculated by the system both on the items and on the whole test. Unfortunately, the four rules on which the system is based seem to be insufficient to cover all of the possible defects which can affect an item. Moreover, these rules are not inferred from a solid knowledge-base and use crisp values (i.e., one of them, states that an option must be discarded if its frequency is 0, independently from the size of the sample). Furthermore, it does not contain any experiment which demonstrates the effectiveness of the system in improving assessment. Nevertheless, this work has given us many ideas, and our work can be considered a continuation of it.

7. CONCLUSION

In this paper we have presented an e-testing system, capable of improving the overall quality of the items used by the tutors, through the re-use of the items across subsequent on-line test sessions. Our system's rules use statistical indicators from the IA model to measure item quality, to detect anomalies on the items, and to give advise for their improvement. Obviously, the system can only detect defects which are visible analyzing results of item and distractor analysis indicators.

The strength of our system is in the possibility for all the tutors, and not only experts of assessment or statistics, to improve test quality, by discarding or, where possible, by modifying defective items. The system has been used at the University of Salerno, to assess the students of a course. This initial experiment has produced encouraging results, showing that the system can effectively help the tutors to obtain items which better discriminate between strong and weak students and better match the difficulty estimated by the tutor. More accurate experiments, involving a larger set of items and students, are necessary to effectively measure the system capabilities.

Our system performs a classification of items, carried out by evaluating fuzzy rules. At present, we are collecting data on test outcomes. Once a large database of items and learner's answers is available, there will be the possibility of exploiting other methods of classification, based on data, such as hierarchical methods, K-means methods, and correlation methods.

8. REFERENCES

[1]. Bardossy, A., Duckstein, L., *Fuzzy Rule-Based Modeling with Applications to Geophysical, Biological, and Engineering Systems*. CRC Press. 1995

[2]. *Blackboard*. http://www.blackboard.com/, 2006

[3]. Chen, C.M., Duh, L.J., Liu, C.Y. A Personalized Courseware Recommendation System Based on Fuzzy Item Response Theory. *In Proceedings of IEEE Int. Conf. on e-Technology, e-Commerce and e-Service*, 2004, Taipei, Taiwan, pp. 305-308

[4]. Civanlar, M. R., Trussel, H. J. Constructing membership functions using statistical data. *Fuzzy Sets and Systems*. 18. 1986, pp. 1-14

[5]. Costagliola, G., Ferrucci, F., Fuccella, V., Gioviale, F., A Web-based Tool for Assessment and Self-Assessment, *in Proceedings of 2nd International Conference on Information Technology: Research and Education*, ITRE'04, pp. 131-135

[6]. *ECMAScript*. Standard ECMA-262, ECMAScript Language Specification, http://www.ecma-international.org /publications/files/ECMA-ST/Ecma-262.pdf, 2005

[7]. FCL. *Fuzzy Control Prog. Committee Draft CD 1.0* (Rel. 19 Jan 97). http://www.fuzzytech.com/binaries/ieccd1.pdf

[8]. Hambleton R. K., Swaminathan, H. *Item Response Theory-- Principles und Applications*, Netherlands: Kluwer Academic Publishers Group, 1985.

[9]. *Hibernate*. http://www.hibernate.org, 2006

[10]. *jFuzzyLogic*. Open Source Fuzzy Logic (Java). http://jfuzzylogic.sourceforge.net/html/index.html, 2006

[11]. Ho, R.G., Yen, Y.C. Design and Evaluation of an XML-Based Platform-Independent Computerized Adaptive TestingSystem. *IEEE Transactions on Education*, 2005, VOL. 48, NO. 2, , pp. 230-237

[12]. Hsieh, C.T., Shih, T.K., Chang, W.C., Ko, W.C. Feedback and Analysis from Assessment Metadata in E-learning. *In Proceedings of 17th International Conference on Advanced Information Networking and Applications*, 2003, Xi'an, China, pp. 155-158

[13]. Hung, J.C., Lin, L.J., Chang, W.C., Shih, T.K., Hsu, H.H., Chang, H.B., Chang, H.P., Huang, K.H., A Cognition Assessment Authoring System for E-Learning. *In Proc. of 24th Int. Conf. on Distributed Computing Systems Workshops* '04, pp. 262-267

[14]. Johnson, S.C. Hierarchical Clustering Schemes. *Psychometrika* 32, 1967, pp. 241-254.

[15]. Lloyd, S. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*. 28, 2. pp. 129 – 137, 1982

[16]. Massey. *The Relationship Between the Popularity of Questions and Their Difficulty Level in Examinations Which Allow a Choice of Question.* Occasional Publication of The Test Dev. and Res. Unit, Cambridge.

[17]. *Moodle*. http://moodle.org/, 2006

[18]. *Questionmark*. http://www.questionmark.com/, 2006

[19]. Schwartz, M., Task Force on Bias-Free Language. *Guidelines for Bias-Free Writing.* Indiana University Press, Bloomington, IN, 1995

[20]. Stage, C. *A Comparison Between Item Analysis Based on Item Response Theory and Classical Test Theory*. A Study of the SweSAT Subtest READ, http:// www. umu. se/ edmeas/ publikationer/ pdf/ enr3098sec.pdf, 1999

[21]. Sun, K. T., An Effective Item Selection Method for Educational Measurement, *In Proceedings of International Workshop on Advanced Learning Technologies*, 2000, pp. 105-106

[22]. Woodford, K., Bancroft, P. Multiple Choice Items Not Considered Harmful, *in Proceedings of the 7th Australian conference on Computing education*, Newcastle, Australia, 2005, pp. 109-116

[23]. Zadeh, L. A. *Fuzzy sets and their applications to classification and clustering.* Academic Press (1977).