

## A Two-Stage Multi-Agent Based Assessment Approach to Enhance Students' Learning Motivation through Negotiated Skills Assessment

**Abdelhafid Chadli, Fatima Bendella and Erwan Tranvouez**

Computer Science Department, Ibn khaldoun University of Tiaret, Algeria // Computer Science Department, University of Sciences and Technology of Oran, Oran, Algeria // Data, information & content management group, Aix Marseille University, CNRS, LSIS, UMR 7296, Marseilles, France // chadli.abdelhafith@univ-tiaret.dz // fatima.bendella@univ-usto.dz // erwan.tranvouez@lsis.org

### ABSTRACT

In this paper we present an Agent-based evaluation approach in a context of Multi-agent simulation learning systems. Our evaluation model is based on a two stage assessment approach: (1) a Distributed skill evaluation combining agents and fuzzy sets theory; and (2) a Negotiation based evaluation of students' performance during a training simulation or a Problem-solving process in a Computer-assisted learning system. This paper highlights how this approach deals with the problem of subjective evaluation of students, and shows the impact of Negotiated skills evaluation on reducing the students' rate of dropout. This approach can also compensate the absence of human expert for assessing training results. Applied to training in plant protection, experiments' results showed first the fuzzy sets based assessment to be similar to the domain expert's assessment and second the negotiated skills assessment to be effective in assessing students' abilities and sustaining students' motivation to continue learning. This evaluation approach allows us to address the problem of subjective assessment and overcome some difficulties encountered in traditional measurement models.

### Keywords

Agent-based evaluation, Negotiated collaborative evaluation, Distributed evaluation, Fuzzy logic based assessment

### Introduction

The use of Simulation based systems for education and training purposes is still hindered by lack of methods and tools to assess learners' progress during a training session. For instance, in classroom-based learning, assessment is usually conducted in two ways (*formative* and *summative*) and is performed by human experts. However, in Simulation-based learning, these assessment methods become inappropriate, as they often consist in a negative feedback without explanation or improvement guidance, which can lead a learner to lose motivation and to stop learning. Furthermore, when it comes to assessment, there is no appropriate Computer-based assessment methodology adapted to Simulation-based learning/training (Ekanayake et al., 2011). Currently, skills assessment in training simulations is often conducted by human instructors using subjective qualitative methods (based on human expertise), which becomes difficult to automate as expected in Simulation-based learning systems in regards to reduction of instructional time and costs (Eck, 2006).

In order to help students better cope with difficulties encountered in solving problems, many researchers have developed intelligent assessment tools based on artificial intelligence approaches (Stathacopoulo, 2005; Huang, 2008). For example, the conceptual framework developed by Mislevy et al. (2003) adopts an Evidence-Centered Design (ECD), which informs the design of valid assessments and can yield real-time estimates of students' competency levels across a range of knowledge and skills. However, the following issues in existing assessment models require further investigation:

- Assessment tools often proceed in a single stage evaluation of student's skills, and focus more on producing marks than giving detailed explanation on what the students failed to understand or put into practice (Chang et al., 2006). Furthermore, learner feedbacks may be insufficient and lack accuracy to help students.
- Generally, all assessment tools set a threshold score for tests to be succeeded. This discriminates students whose final score is near the passing limit. Does a student with a 9.9 final score have significantly less knowledge than a student with 10 as final mark? Moreover when considering potential error margins.
- The existing assessment tools focus more on assessing learner's performance regardless of whether this assessment contributes to learners' motivation and not give up of learning.

These issues can be addressed first by refining the assessment skills criteria in order to detail what part of the learning process went wrong. Secondly, marks should be handled with a margin error, thus avoiding threshold phenomena where an assessment can change significantly. Moreover, taking into account limited compensation

between the different assessment criteria (as when deciding whether a student should graduate or not), would grant a more flexible assessment as humans do. Finally, reporting feedback to the student can then be detailed and not fully negative, reducing the demotivation issue.

In this paper, we propose to use distributed assessor agents to assess skills individually and thus inform precisely about difficulties encountered by the student at each skill. By using Fuzzy sets, assessor agents are able to evaluate the level of control of each skill by considering the difficulty of each action of the skill. Our strategy involves a two-stage approach (see figure 1): The first stage focuses on student's skill evaluation by means of assessor agents; each is responsible of evaluating only one skill of the student. This will inform the second stage of the approach, which concerns the global evaluation of the student's capabilities. This evaluation stage is managed by an aggregate agent and is based on the assessor agents' assessments, allowing a negotiation process to decide whether the student passes the required skills qualification.

The proposed system was evaluated by conducting three experiments using students training in plant protection as subjects. The following three issues are explored here:

- Comparing the students' assessment by the system to an expert like.
- Whether provided feedback helped students with their problem solving.
- Whether assessment method encouraged students to continue learning and not give up the learning system sessions.

This paper is structured in four parts. We first present relevant research works in learning assessment area, related issues and the problematic we address. Then we describe our skills evaluation approach, in this part we detail firstly the skill evaluation and secondly the global evaluation where we expose all related aspects to the negotiation process. Finally, we present the experimental design and analytical results before drawing the conclusion.

## **Relevant research works and related issues**

A number of intelligent learning environments are concerned with evaluating learners in a purpose of providing individual learning guidance to learners, and enhancing the performance of each learner. For example, Stathacopoulou et al. (2005) have developed a neural network implementation for a Fuzzy logic-based model of the diagnostic process as a means to achieve accurate student diagnosis in intelligent learning environments. Huang et al. (2008) have proposed an intelligent diagnosis and assessment tool, and have incorporated it into an open software e-learning platform developed for programming language courses, based on text mining and Machine learning techniques to alleviate teachers' workload. In order to address the question: *How can Game-based learning be assessed?* Shute and Ke (2012) have adopted a form of stealth assessment based on ECD-based assessments. All these works led to powerful tools that circumvent the weaknesses of the traditional assessment. However, these assessment tools did not deal with the issues referred to above.

Distributed skill centered assessment is a method that can address these problems by combining heterogeneous skill assessment methods and thus enable the development of robust and valid simulation or Problem solving based learning systems (Oulhaci et al., 2013). Additionally, the presence of imperfect information is an important factor which often leads to errors in the learner evaluation. As Fuzzy sets theory is classically used to handle uncertainty by using qualitative output variable, using the learner's level over a knowledge classification scale can help solving this limitation. Many studies focusing on assessment have used the Fuzzy theory, for example, Innocent et al. (2005) have developed a Fuzzy expert system for the medical diagnosis to show how fuzzy representations are useful for taking uncertainty into account and can be applied to model the acquisition of knowledge by experts. In our assessment tool, the first stage's assessor agents are in charge of applying the Fuzzy logic process to evaluate each skill master degree.

In this research, we also promote the use of Negotiated collaborative assessment based on Fuzzy sets individual evaluation. Some researches proved that discussion and negotiation between independent assessors can enhance the reliability of assessment criteria for portfolios (Rees & Sheard, 2004). According to Pitts et al. (2002), the collaborative assessment makes it possible to provide an enhanced assessment of learner so as to improve his skills. This idea was used first in open model of the student's knowledge (Dimitrova et al., 1999), which involved the student and the assessor (or an agent acting on the assessor's behalf) negotiating an agreed assessment of the student

(Brna et al., 1999). Negotiation is also used in peer assessment, Lai and Lan (2006) have developed a method for modeling collaborative learning as multi-issue agent negotiation using fuzzy constraints. The proposed method aggregates student marks to reduce personal bias. In this framework, students define individual Fuzzy membership functions based on their evaluation concepts and agents facilitate student-student negotiations during the assessment process. In the same context, several authors show that the allocation of bonuses to students is a relevant intervention strategy for cognitive engagement and student achievement (Black & Duhon, 2003; Tuckman, 1996).

The approach that we propose allows learners to engage in a Simulation based learning and offers the opportunity for the learner, whose final score is close to the threshold fixed for test success, to negotiate his final score by means of assessor agents through the use of the results of this learner's skills assessments and bonus attribution according to the student's skill performance.

Agents have been widely used in training environments for different purposes. Multi agent modeling is probably the most used tool in these environments, notably by using their intelligence ability and autonomy for learner's assessment. For example, Baghera is an intelligent tutoring system which uses a theorem prover agent for automatic verification of proofs (Caferri et al., 2000). Pilato et al. (2008) have used a conversational agent to assess the student knowledge through a natural language question/answer procedure. They have used the latent semantic analysis technique to determine the correctness of the student sentences in order to establish which concepts the student knows. An assessment agent is used by Lai and Lan (2006), this approach allows student, whose coursework is marked, negotiate with markers using this assessment agent to reach a final assessment.

In conclusion, the discussion above reveals that combining assessor agents, Fuzzy sets theory and negotiation in an assessment tool is an idea that deserves investigation.

## The evaluation modeling

In this section, we present our two-stage evaluation approach, from the Fuzzy logic modeling of the individual skill assessment to the negotiation process taking place between assessor agents and mediated by the aggregate agent (see figure. 1). Our approach follows three steps:

### Step1: Skills identification step

In order to better identify and to characterize the required skills within any of the domains of competence, we propose to follow the knowledge classification classically used in metacognition theory (Schraw & Moshman, 1995), where three types of knowledge are defined: *Procedural knowledge*, *Declarative knowledge* and *Conditional knowledge*.

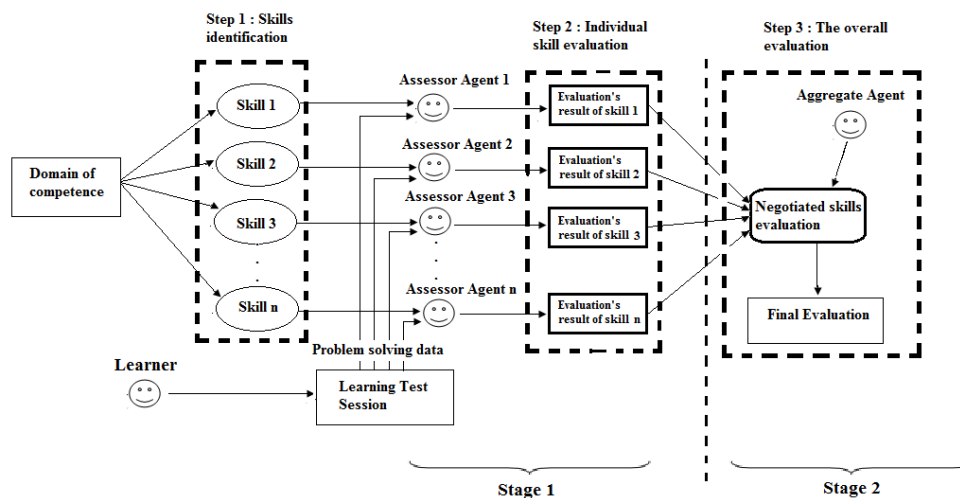


Figure 1. General overview of the system

Using these types of knowledge analysis, the main competences of the domain are first identified and are later divided in sub competences using a goal oriented analysis (see figure 2), until reaching the actions to be performed by the learner via the learning system interfaces. For example in the domain of Word-based mathematical problem solving, four main skills are identified: 1-understanding the problem, 2-making the plan of resolution, 3-executing the plan and 4-reviewing the solution. These competencies are considered as main skills that student must master to solve a Word-based mathematical problem. Then other skills are derived from the main ones until obtaining actions, reflecting identified skills, that student can carry out via the Computer based learning system.

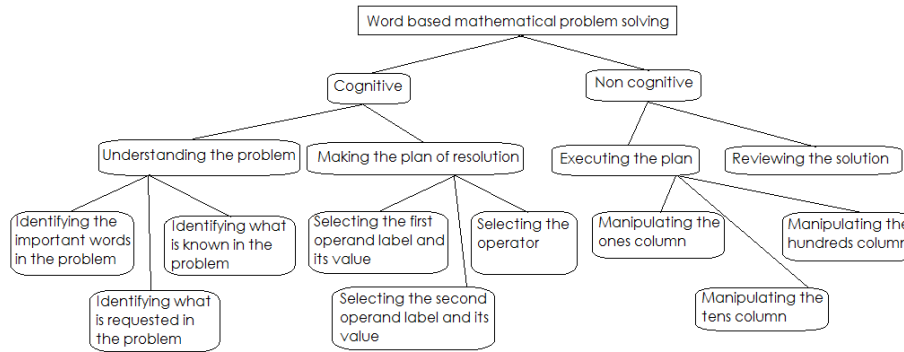


Figure 2. Illustration of a competency model for word based mathematical problem solving

### Step 2: The step of skill evaluation

In our evaluation approach, each skill within a domain of competence is represented by one or more actions that learner can perform, each action's difficulty being qualitatively measured. We have defined according to Problem based learning experts, three levels of difficulty: (1) Actions with high level of difficulty, (2) Actions with average level of difficulty and (3) Actions with weak level of difficulty. Moreover, the scale used for the classification of learners' competency control is inspired by the traditional classification of the performances employed by the assessors: (1) insufficient, (2) acceptable and (3) satisfactory (Merrill, 1983). For example, in the domain of Word based mathematical problem solving, Polya (1945) has identified four competences. In our system, each of them will be evaluated by an assessor agent through the evaluation of the learner's performed actions (e.g., the competence called *understanding the problem* is characterized by many actions as shown in Table 1). To produce individual action evaluation, a fuzzy model is defined.

Table 1. Illustration of actions to be performed

| Action   | Level of difficulty |
|--|---------------------|
| Identifying important words in the problem               | High                |
| Identifying what is known in the problem                 | Average             |
| Identifying what is requested in the problem             | Average             |
| Defining the position of the missing part of the problem | Weak                |

The fuzzy model represents the expert's knowledge in linguistic form and includes the characteristics of the learner in the form of a set of fuzzy systems thus allowing an evaluation similar to that of an expert. A Fuzzy set is characterized by a membership function  $f: E \rightarrow [0,1]$ , which positions the members of the universe of discourse E in the units interval [0,1] (Zadeh, 1965). Value 1 (*resp.* 0) means that the member is (*resp.* not) included in the given set. The values between 0 and 1 characterize the fuzzy members. In our case, the universe of discourse E corresponds to the percentage of the actions performed correctly (classified by type and weighted by level of difficulty), and is divided in 11 elements {0,10,20,30,40,50,60,70,80,90,100}. For the transformation phase of quantitative variables into Fuzzy logical variables (i.e., Fuzzification process), we have defined four variables for each category of action (i.e., level of difficulty):

- Three input variables  $VE_n$  ( $n \in \{1,2,3\}$ ) (one for each level of difficulty) and
- One output variable VS representing the learner's qualitative level of knowledge or mastership of one skill {Very weak, Weak, Insufficient, Average, Good, Very good, Excellent}.

For each one of these variables a membership function is defined in collaboration with experts in education (see figure 3 example). All assessor agents will use inference rules, based on these membership functions, to position the learner's level on the classification scale. We have established 27 inference rules (3 levels of difficulty and 3 classification values for each level) corresponding to the various evaluations related to the three levels of difficulty (see figure 4 example).

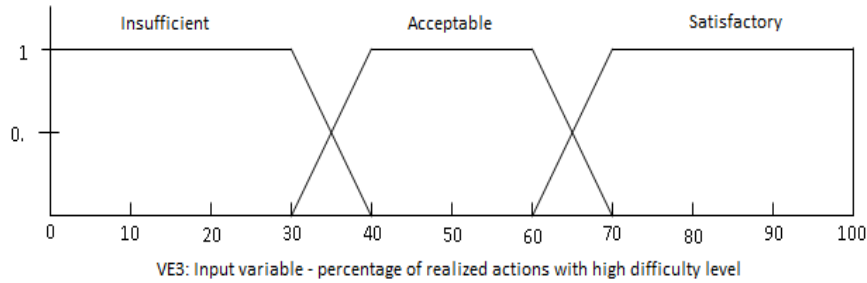


Figure 3. The membership function for the input variable VE3

|  |
|--|
| <p>If (EAWD = Satisfactory) AND (EAAD = Satisfactory) AND (EAHD = Insufficient)<br/> Then (Knowledge level VS = Average)</p> <p>Where EAXD = Evaluation of actions with X level of difficulty<br/> { X from W: Weak, A: Average, H: High }</p> |
|--|

Figure 4. Example of an inference rule

Once computed, this fuzzy information must be converted into a real estimated value (i.e., defuzzified) which will represent the judgment of the assessor agent. The Defuzzification method produces a VS value as a number rounded to the nearest integer between 1 and 7 (1,2,3,4,5,6,7) corresponding respectively to the 7 evaluation levels (Very weak, Weak, Insufficient, Average, Good, Very good, Excellent) . At the end of the learning test session, each assessor agent has estimated the learner's mastership level in "its" skill. This first stage evaluation will contribute to the overall evaluation of the second stage directed by the aggregate agent.

### Step 3: The overall evaluation step

Generally, tutors set a threshold score for tests to be succeeded, for example in classrooms this score is fixed at 10 (i.e., half of full score). In our approach, we propose that each skill is considered mastered if the learner obtains a score equal or upper the value 5 (i.e., VS = 5, GOOD). However, we tolerate compensation between different skills with a minimum threshold score for each skill; unlike other domain of knowledge such as medicine where compensation is not accepted (Frank et al., 2010). In skills diagnosis, two classes of models have been commonly used: conjunctive models and compensatory models. When a domain of knowledge involves multiple skills and when the low mastery of a single one of them is sufficient for failing this domain knowledge, the model is considered part of the *conjunctive class*, meaning that all skills are necessary. Conversely, if a strong mastery of some skills is sufficient to succeed a test, it will be considered part of the *compensatory class* of models (Desmarais et al., 2012; Roussos et al., 2007). Our approach combines both model classes, by defining a minimum threshold degree of mastership of each skill and allowing compensation between skills by means of bonuses attribution for high level skill mastery.

At the reception of all evaluations (provided by assessor agents), the aggregate agent analyzes individually each one of them and computes the average score if necessary. Three decisions can be taken: the learner passes or fails the test, or the learner assessment needs negotiation. Analysis of the aggregate agent follows the algorithm shown in Figure 5.

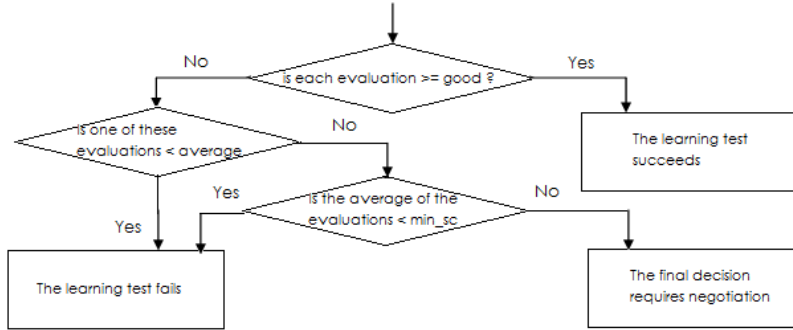


Figure 5. The overall evaluation decision

Bonus marks benefits can improve students' attitude or effort, completion of work on time and/or neatness of finished work. In our approach, we propose to award bonuses to encourage the learner according to his skills performance. Obtaining a bonus allows to decrease the test validation score for a learner (initially equal to 5) of value " $vb1$ " (i.e., value to be defined by domain of competence experts) for the first bonus and of " $vb2$ " for the second bonus, this score diminution is used in the negotiation process to help learners whose global score is close to the validation score. For example, if a learner's averaged score is equal to 4.66 (i.e., lower than the preset validation score,  $4.66 < 5$ ) and an assessor agent awarded him *bonus1* and *bonus2*, the new validation score to be proposed by this assessor agent during the negotiation process for this learner is  $5 - (vb1 + vb2)$ .

#### Computation details on how bonus marks guide the negotiation

In our approach, three kinds of bonuses are awarded by agents according to student's performance:

**Bonus 1:** each assessor agent decides on the attribution of *bonus1* depending on the score obtained in the skill performance.

- **IF (skill's score  $\geq 6$ )** then the assessor agent allots *bonus1* to the learner (reduction in the validation score of a value of  $vb1$ , the agents accepts to validate the learner's test with a score of  $5 - vb1$ ),
- **IF (skill's score  $\in [5, 6[$ )** then he allots half of *bonus1* to the learner (reduction in the validation score of a value of  $(vb1)/2$ ),
- **IF (skill's score  $< 5$ )** then he don't allot *bonus1* to the learner.

**Bonus 2:** is allotted by the assessor agent if the learner performs correctly at least 80% of actions with high level of difficulty, otherwise the learner does not receive this bonus. Obtaining *bonus2* allows to decrease the global validation score of a value of  $vb2$  (i.e., the assessor agent accepts to validate the learner's test with a score of  $5 - vb2$ ).

**Bonus 3:** the aggregate agent can allot one third bonus which consists in improving the learner's score by a value of  $vb3$  if this latter succeeded in at least  $2/3$  of all skills (e.g., if the learner's averaged score of all skills is 4.66, the aggregate agent proposes to validate the learner's test with a new score equal to  $4.66 + vb3$ ).

Once bonuses are calculated, each assessor agent builds its satisfaction table and assigns a satisfaction degree to each validation score that he can propose to the aggregate agent during the negotiations (i.e., the satisfaction degree is used by the assessor agent as an indicator to propose or not a new validation score). This satisfaction table will be used in the process of negotiation; the assessor agent begins by proposing the first validation value in its table and continues until an agreement is reached or the actual degree of satisfaction is null (i.e., the assessor agent rejects the negotiation). An example of four satisfaction tables corresponding to four different cases of bonuses attribution is presented in table 2 (with  $vb1 = 0.34$  and  $vb2 = 0.17$ ) where no bonuses are attributed in case (a), only *bonus2* is given in case (b),  $\frac{1}{2}$  *bonus1* + *bonus2* are attributed in case (c) and both of bonuses 1 and 2 are awarded in case (d).

For example in case (c), the first validation value is set to 4.83 (i.e.,  $4.83 = 5 - 0.17$ ) after attribution of  $\frac{1}{2}$  *bonus1*, the degree of satisfaction corresponding to this value is 2 because there is another validation value: 4.66 (i.e.,  $4.66 = 4.83 - 0.17$ ) computed after awarding of *bonus2*. The degree of satisfaction of this second validation value is 1 since there are no more proposals. The last value 4.5 has a null degree of satisfaction; this means that if a proposal of test

validation score is equal or less than 4.50, the assessor agent will reject this proposal. Note that the higher the degree of satisfaction is, the higher is the belief of the assessor agent to propose the corresponding validation value.

Table 2. Four examples of satisfaction tables' construction

| Case | Bonus | Validation Value | Degree of satisfaction |
|------|-------|------------------|------------------------|
| A    | 0     | 5.00             | 1                      |
|      | 0     | 4.83             | 0                      |
| B    | 2     | 4.83             | 1                      |
|      | 0     | 4.66             | 0                      |
| C    | 1     | 4.83             | 2                      |
|      | 2     | 4.66             | 1                      |
|      | 0     | 4.50             | 0                      |
| D    | 1     | 4.83             | 3                      |
|      | 1     | 4.66             | 2                      |
|      | 2     | 4.50             | 1                      |
|      | 0     | 4.33             | 0                      |

### From evaluation marks to agents negotiation

The implementation of the negotiation needs the definition of a communication protocol between agents and the development of private strategies for each agent (Jennings et al., 2001). We present here the negotiation protocol used in our evaluation system and the rules employed by agents to reach a compromise. The objective of the protocol is to define the messages exchanged by the agents. The negotiation protocol that we propose is characterized by a succession of messages exchanged between an initiator (aggregate agent) and participants (assessor agents) as in Contract Net Protocol (Smith, 1980). We introduced a new idea in the negotiation process which consists in making a promise of concession by the aggregate agent if the assessor agent agrees to make a concession in turn (i.e., concession = one bonus consumption).

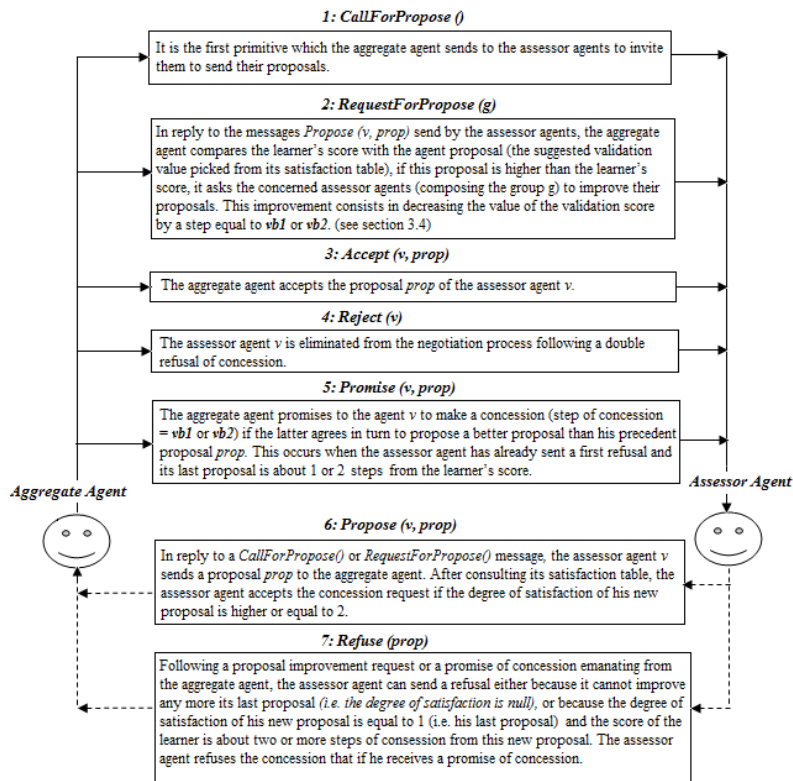


Figure 6. Exchanging messages between the aggregate agent and the assessor agents

Figure 6 describes the conversation between the aggregate agent and assessor agents. The aggregate agent starts by sending a *CallForProposal* message to the assessor agents, the parameter “g” indicates the group of concerned agents. All assessor agents reply by sending their proposals. The negotiation continues until aggregate agent sends an accept message to at least 2/3 of assessor agents (i.e., the test succeeds with a score equal to the last proposal done by assessor agents) or if the aggregate agent removes more than half of assessor agents from the negotiation process (i.e., the learner’s test failed).

## Experiments and results

In order to verify the effectiveness of our assessment approach and to validate its performance, we have chosen the domain of competence of plant protection where a simulation crop pest control training environment is already developed to improve training of future phytosanitary agents. In this learning system, the learners’ behaviors have to be observed and analyzed in term of expected skills. However the systematic presence of a human expert cannot be ensured requiring some Computer-assisted assessment. The experts of this domain of competence have identified three main skills:

- *Control Skill*: as the ability of the learner to identify and to exploit several indicators (for ex. estimating a proliferation risk requires knowing current meteorological data, past data on pests’ location ...).
- *Intervention skill*: representing the achievement of one or several tasks correctly sequenced and accomplished at the right time.
- *Assessment skill*: the ability of the learner to assess a situation or a resources requirement considering scenario’s information and to act accordingly. It thus precedes one or several actions of intervention.

Following our two-stage evaluation process, we have developed three assessor agents and an aggregate agent. All agents are implemented in Jade (Java Agent Development Environment). Figure 7 displays messages exchanged between agents during the negotiation process and the final result is announced by the aggregate agent. Thresholds values (minimum test success score, maximum test failure score, *bonus1*, *bonus2* and *bonus3* values) were established by plant protection experts. The experiments we have conducted involved fifty students training in plant protection - zoophytia. These students were divided randomly into two groups: A (experiment group) and B (control group) of twenty five students each. To verify the effectiveness of our assessment system and to validate its performance, we conducted three experiments. Microsoft Excel was used for all data manipulation and statistical tests.

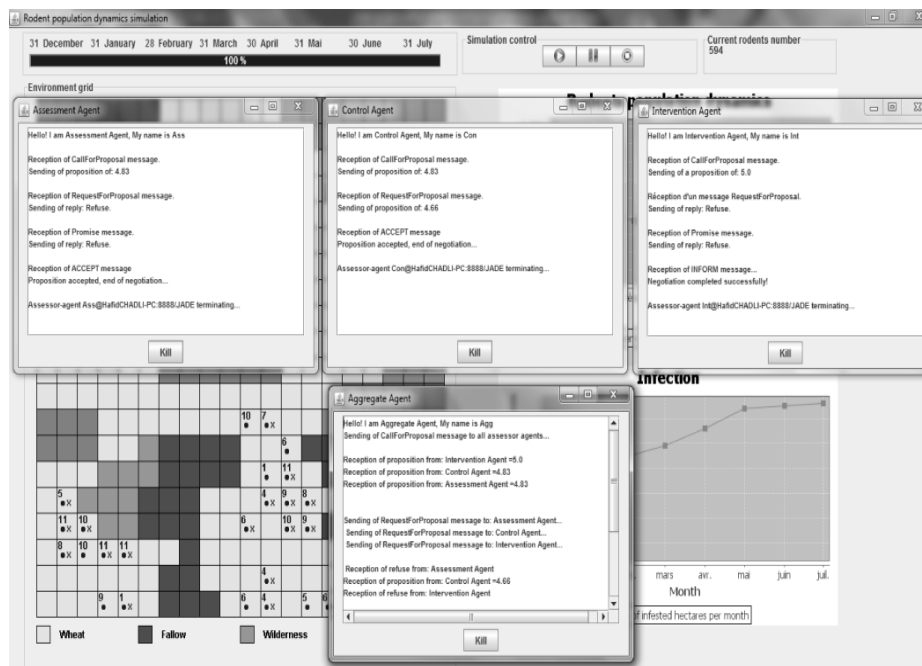


Figure 7. The screenshot of the overall evaluation via negotiation between the assessors



## First experiment

The objective of this experiment is to test the effectiveness of the evaluation system, precisely, the first stage evaluation involving only assessor agents. To this end, students in group A were assessed during two learning test sessions by the system (i.e., our three assessor agents), and after adaptation of the same test scenario on paper, the responses of these same students were again assessed by an expert of the domain taking into account the number, type and difficulty of the actions carried out correctly. In order to avoid bias between the system's evaluation and that of the expert one, we intentionally conducted two learning test sessions. We adopt the Pearson correlation to compare the correlations between scores given by the system (assessor agents' scores) and scores of the expert for each type of skill and for both sessions (i.e., all scores are between 1 and 7, see section 3.2). Table 3 illustrates this experiment's results.

Table 3. Correlation analysis between system and expert assessments

| Type of competence | Session 1<br>Correlation | Critical<br>value | Session 2<br>Correlation | Critical<br>value | Level of<br>Significance |
|--------------------|--------------------------|-------------------|--------------------------|-------------------|--------------------------|
| Control skill      | 0.7618                   | 0.396             | 0.8991                   | 0.396             | 0.05                     |
| Intervention skill | 0.8111                   |                   | 0.8326                   |                   |                          |
| Assessment skill   | 0.8101                   |                   | 0.7144                   |                   |                          |

**Discussion:** This first experiment intended to examine the effectiveness of our assessment system. The best way to achieve this is to compare the scores assigned by the system to those given by the domain expert during two rounds. Significant correlation (correlation is greater than the critical value for the two sessions) between the two evaluations in this experiment constitutes a preliminary validation of the effectiveness of our assessment system. We are however aware of the limitations of our validation since our approach is conducted in only one simulation learning system. Also, the results validate the claim regarding the importance of embedding Fuzzy sets in assessment process. By providing actions with several levels of difficulty, the assessment system is able to accurately point the learner's level of knowledge with a certain granularity for each competence of the domain. This has some influence in the training system software requirements.

Furthermore, dedicating specialized agent for assessing each skill individually has several advantages: The diversity of agents enables a distributed assessment and ease of expandability of assessed skills. Heterogeneous agents can thus evaluate specific information (learner's answers versus predefined correct answers, simulation results, users' interactions with the training system etc.). As such, the use of agents allows defining independently each skill assessment method (different criteria, parameters and algorithms).

## Second experiment

In order to test the performance of students evaluated by the system, we have organized two learning test sessions where students in Group A are evaluated by the system and receive feedback at the end of the first learning test session allowing them to revise their behaviors. For the same test scenario, students in Group B are evaluated by the domain expert and receive no explanation on their behaviors. The comparison between behaviors during the first session and those of the second session, for both groups, allows to rule on the performance of the students. We propose a paired t-test analysis to verify the improvement of behaviors in each group. Table 4 represents the analysis of the performances for the group A and the group B.

Table 4. Performance analysis

| Type of competence | Group A |         |                   |                   | Group B |         |                   |                   |
|--------------------|---------|---------|-------------------|-------------------|---------|---------|-------------------|-------------------|
|                    | t-value | p-value | Session 1<br>mean | Session 2<br>mean | t-value | p-value | Session 1<br>mean | Session 2<br>mean |
| Control skill      | 2.6134  | 0.0152  | 3.92              | 4.40              | 1.6164  | 0.1191  | 3.52              | 3.16              |
| Intervention skill | 2.3842  | 0.0254  | 3.52              | 4.12              | 1.7321  | 0.0961  | 3.40              | 3.00              |
| Assessment skill   | 2.5981  | 0.0158  | 3.32              | 3.92              | 2.00    | 0.0569  | 3.44              | 3.84              |

**Discussion:** Regarding the performance of students evaluated by the system, we have conducted two learning test sessions. Comparison of students' scores obtained during the first and second learning test session and compared to those of the first session indicates whether or not learners have improved their behaviors. Improvement of behavior between the two sessions confirms the accuracy of assessment and the quality of feedback. As all p-values (group A's experiment) don't exceed the level of significance  $\alpha$  (with  $\alpha = 0.05$ ) we conclude that the difference between the two groups was significant. Analytical results indicate that students of group A had acquired more knowledge than students of group B since the means of scores for two types of competence (control and intervention) increased significantly from 3.92 and 3.52 to respectively 4.40 and 4.12. However, for the assessment skill, means of scores gained a slight improvement from 3.32 to 3.92; we believe this is due to the assessing complexity of some situations in the simulation learning environment. Furthermore, means of scores of students in group B remained constant except for assessment skill where scores have risen moderately. All these results point to the conclusion that learners learned from the feedback generated by the assessment system.

Most students of group A confirm that the system helped them to review and improve their behaviors. They also felt that the flexibility of the assessment through bonuses' awarding and negotiation is a motivating factor that prevents loss of motivation and give up of the learning system besides improving the overall behavior of the learner. However, some would think that if students are involved in the evaluation they could argue their choice of actions. In summary, assessing learners differently than existing approaches, which assessment frequently focus on the result regardless the learner's skills involved to obtain this result, allows us to address the problem of subjective assessment and overcome some difficulties encountered in traditional measurement models. Our approach can support learners to build problem-solving skills, and to enhance their behaviors over learning test sessions through independent assessor agents.

### Third experiment

In order to measure the impact of the Negotiated assessment on student's motivation to continue learning and not give up the learning system, we have organized six learning test sessions for both groups (2 sessions per day). Students of group A were assessed using both stages of the evaluation system (i.e., individual skills assessments and the negotiation mechanism) while students of group B were evaluated only by the first stage of our evaluation model without any possibility of negotiation between assessor agents (if the student's score  $\geq 5$  then the test is succeeded else it failed). Specifically, we focus on the participation rate and the dropout rate of both groups' students having obtained consecutively during many sessions a score in the interval  $[4.33, 5[$  (i.e.,  $min\_sc = 4.33$ , this value was fixed by the domain experts, it allows negotiation between assessor agents if the score of a learner in group A is  $> 4.33$  and punishes a student of group B by a failure of test if his score is  $< 5$  ). Note that for this experiment, we do not consider the abandonment of learners with other scores. The students were informed that they could leave the learning test session at any time and give reasons for their abandonment by mean of the questionnaire provided at each session. Table 5 shows respectively for each group the number of learners having obtained successively for  $x$  times a score in the interval  $[4.33, 5[$  (with  $x \in [2,6]$ ), followed by the number of learners among them who have succeeded their learning test at this attempt (i.e., by mean of negotiation), followed by dropout rate of this attempt. All 50 students attended the session 1.

Table 5. Participation and dropout rates

| Attempt | X = 2  | X = 3  | X = 4 | X = 5 | X = 6 | Total of dropouts |
|---------|--------|--------|-------|-------|-------|-------------------|
| Group A | 14/7/0 | 10/7/0 | 9/4/0 | 8/5/1 | 5/4/1 | 2                 |
| Group B | 9/0/0  | 7/0/1  | 5/0/3 | 2/0/2 | 0/0/0 | 6                 |

**Discussion:** We note that the dropout rate in group B was higher than that in group A for that category of learners although it is not very high (i.e., 6 Vs 2). In order to identify the reasons of this abandonment, we have analyzed the answers of targeted learners to the questionnaire provided for indicating the reasons of giving up the learning test session (see Table 6). In addition, the low dropout rate of group A's learners is due to the high success rate of learners unlike learners of group B by dint of the flexibility of the negotiated assessment as shown in Table 5.

Table 6. Learners' answers about giving up the learning test session

| Question: Why you no longer want to continue learning tests?                   | Group A | Group B |
|--|---------|---------|
| The test is difficult  | 1       | 1       |
| The evaluation is subjective   | 1       | 5       |
| The learning environment is not interesting                                    | 0       | 0       |
| The simulation learning environment does not reflect the reality on the ground | 0       | 0       |
| Other  | 0       | 0       |

All answers (except one) of group B's dropouts (5 of 6) for this category of learners point to the conclusion that the evaluation is subjective, students report that they deserve to succeed their learning tests sessions at many times while the system considered they have failed, it is precisely for this reason that we chose this category of learners (close to success, but unfortunately due to a rigid common evaluation, the system decides that they failed their tests). While in group A, only one learner having forsaken the learning test session considered that the evaluation is subjective. This comparison points to the conclusion that the main cause of the abandonment of learners is due to the subjectivity of the adopted assessment mechanism, while the use of the Negotiated evaluation for learners of group A has contributed to the preservation of motivation to continue learning. In addition, we took no heed of dropouts of learners with different consecutive scores (not all belonging to the interval [4.33, 5[ ) so as not to bias the results of the experiment. Although we are aware of the reduced sample of this experiment, we believe that a large-scale experiment will confirm our claim.

### Limitations of the present study

- The participation of students was voluntary and we do not claim that the results of the present study can be generalized.
- The improvement of learner's behavior performance can be attributed to the use of the proposed evaluation model or to the extra practice by solving similar pest control issues. Even in the third experiment, the profit for education is that students preserve motivation for learning even if the experiment is based on a partial comparison between the experimental and the control groups. Undoubtedly the educational phenomena are multidimensional and we cannot control all the possible involved variables.

### Conclusion

This study has presented a pragmatic approach of evaluation that allows giving a ruling on the efficiency of the learners in a context of Multi-agent simulation learning systems. This evaluation approach is to be integrated in Simulation-based learning systems, Game-based learning or any Computer assisted problem-solving based learning. Our evaluation model follows three steps: (1) Identification of relevant skills of the domain of knowledge, (2) Evaluation of the learner compared with these skills and (3) Evaluation of the ability of the learner to solve a problem of this domain of competence. In order to assess learners' competences, we have adopted a strategy that involves a two-stage approach based on a collaborative evaluation system. On the first stage, a number of assessor agents are in charge of appreciating the learners' knowledge compared with identified skills of the domain. The recourse to Fuzzy sets theory at this stage allows an evaluation similar to that of an expert. On the second stage, the Aggregate agent ensures an overall evaluation elaborated on the base of the individual evaluations of each assessor agent. Instead of considering an average threshold score which decides of the success and the failure of the learner's learning test, we have preferred a Negotiated collaborative evaluation similar to academic evaluations which promote students according to their results along the learning period. Thus, in some situations, a negotiation process is initiated by the Aggregate agent where each assessor agent uses the learner's results in its dedicated skill in order to negotiate the learner's test success by mean of bonuses awarding. The result of this negotiation represents the final evaluation.

Experimental results indicate that this model provides an assessment similar to that of an expert and significantly improved learners' performance. Furthermore, the Negotiated assessment part of the evaluation model seems promote motivation of learners as demonstrated in third experiment. Based on these results, we conclude that combining Fuzzy sets and agent negotiation has important merits.

Our evaluation system allows initially a skill estimate, this first stage informs us about the strengths and weaknesses of the learner, and thus we provide high precision recommendations to the learners. Therefore the quality of feedback will generate a positive impact on improving the learner's skill performance. In the second time, the global evaluation enables us to conclude about the effectiveness of the learner to solve the problem. Thus, tutors can easily address learners' weaknesses.

Finally, although the proposed approach of assessment has yielded promising results in promoting learning effectiveness and maintaining students' motivation to continue learning, considerable work remains to be done, including the choice of adequate problem corresponding to the learner's profile and based on his skills assessments. Also we'll consider other learner's abilities in the future like group managing, communication, etc in the context of collaborative problem solving.

## References

Black, H. T., & Duhon, D. L. (2003). Evaluating and improving student achievement in business programs: The effective use of standardized assessment tests. *Journal of Education for Business*, 70(2), 90-98.

Brna, P., Self, J. A., Bull, S., & Pain, H. (1999). Negotiated collaborative assessment through collaborative student modelling. In R. Morales, H. Pain, S. Bull, & J. Kay (Eds.), *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 35-42). Le Mans, France: IOS.

Caferra, R., Peltier, N., & Puitg, F. (2000). Emphasizing human techniques in geometry automated theorem proving: A practical realization. In J. Richter-Gebert, & D. Wang (Eds.), *Proceedings of the Workshop on Automated Deduction in Geometry* (pp. 268-305). London, UK: Springer.

Chang, K. E., Sung, Y. T., & Lin, S. F. (2006). Computer-assisted learning for mathematical problem solving. *Computers & Education*, 46(2), 140-151.

Desmarais, M. C., Shaun R., & De Baker, J. (2012). A review of recent advances in learner and skill modeling in intelligent learning environments. *User Modeling and User-Adapted. Interaction*, 22(1-2), 9-38.

Dimitrova, V., Self, J. A., & Brna, P. (1999). The interactive maintenance of open learner models. In S. P. Lajoie, & M. Vivet (Eds.), *Proceedings of the International Conference on Artificial Intelligence in Education* (pp. 405-412). Le Mans, France: IOS Press.

Eck, R. V. (2006). Digital game-based learning: It's not just the digital natives who are restless. *EDUCAUSE Review*, 41(2), 16-30.

Ekanayake, H., Backlund, P., Ziemke, T., Ramberg, R., & Hewagamage, K. (2011). Assessing performance competence in training games. In S. D. Mello, A. Graesser, B. Schuller, & J. C. Martin (Eds.) *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction* (pp. 518-527). Berlin, Germany: Springer.

Frank, J. R., Snell, L. S., Cate, O. T., Holmboe, E. S., Carraccio, C., Swing, S. R.,... Harris, K. A. (2010). Competency-based medical education: Theory to practice. *Medical Teacher*, 32(8), 638-645.

Huang, C. -J., Chen, C. -H., Luo, Y. -C., Chen, H. -X., & Chuang, Y. -T. (2008). Developing an intelligent diagnosis and assessment e-learning tool for introductory programming. *Educational Technology & Society*, 11(4), 139-157.

Innocent, P. R., John, R. I., & Garibaldi, J. M. (2005). Fuzzy methods for medical diagnosis. *Applied Artificial Intelligence*, 19(1), 69-98.

Jennings, N. R., Faratin, P., Lomuscio, A. R., Parsons, S., Sierra, C., & Wooldridge, M. (2001). Automated negotiation: Prospects, method and challenges. *International Journal of Group Decision and Negotiation*, 10(2), 199-215.

Lai, K. R., & Lan, C. H. (2006). Modeling peer assessment as agent negotiation in a computer supported collaborative learning environment. *Journal of Educational Technology & Society*, 9(3), 16-26.

Merrill, M. D. (1983). Component display theory. In C. M. Reigeluth (Eds.), *Educational Technology: Instructional Design Theories and Models* (pp. 279-333). Hillsdale, NJ: Lawrence Erlbaum Associates.

- Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1(1), 3–62.
- Oulhaci, A., Tranvouez, E., Fournier, S., & Espinasse, B. (2013). A multi-agent system for learner assessment in serious games: Application to learning processes in crisis management. In R. Wieringa, S. Nurcan, C. Rolland, & J. -L. Cavarero (Eds.), *Proceedings of the Seventh IEEE International Conference on Research Challenges in Information Science* (pp. 1-12). Paris, France: IEEE computer science.
- Pilato, G., Pirrone, R., & Rizzo, R. (2008). A kst-based system for student tutoring. *Applied Artificial Intelligence*, 22(4), 283-308.
- Pitts, J., Colin, C., Thomas, P., & Smith, F. (2002). Enhancing reliability in portfolio assessment: Discussions between assessors. *Medical Teacher* 24, 197–201.
- Polya, G. (1945). *How to solve it*. Princeton, NJ: Princeton University Press.
- Rees, C., & Sheard, C. (2004). The reliability of assessment criteria for undergraduate medical students' communication skills portfolios: The nottingham experience. *Medical Education*, 38(2), 138-144.
- Roussos, L. A., Templin, J. L., & Henson, R. A. (2007). Skills diagnosis using irt-based latent class models. *Journal of Educational Measurement*, 44, 293-311.
- Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review*, 7(4), 351-371.
- Shute, V. J., & Ke, F. (2012). Games, learning, and assessment. In D. Ifenthaler, D. Eseryel, & X. Ge (Eds.), *Assessment in Game-Based Learning: Foundations, Innovations, and Perspectives* (pp. 43-58). New York, NY: Springer.
- Smith, R. G. (1980). The contract net protocol: High-level communication and control in a distributed problem solver. *IEEE Transactions on computers*, C-29(12), 1104-1113.
- Stathacopoulou, R., Magoulas, G., Grigoriadou, M., & Samarakou, M. (2005). Neuro-fuzzy knowledge processing in intelligent learning environments for improved student diagnosis. *Information Sciences*, 170(2-4), 273-307.
- Tuckman, B. W. (1996). The relative effectiveness of incentive motivation and prescribed learning strategy in improving college students' course performance. *Journal of Experimental Education*, 64, 197-210.
- Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3), 338-353.