Diagnosing Self-Efficacy in Intelligent Tutoring Systems: An Empirical Study

Scott W. McQuiggan and James C. Lester

Department of Computer Science, North Carolina State University, Raleigh, NC 27695
{swmcquig, lester}@ncsu.edu

Abstract. Self-efficacy is an individual's belief about her ability to perform well in a given situation. Because self-efficacious students are effective learners, endowing intelligent tutoring systems with the ability to diagnose self-efficacy could lead to improved pedagogy. Self-efficacy is influenced by (and influences) affective state. Thus, physiological data might be used to predict a students' level of self-efficacy. This paper investigates an inductive approach to automatically constructing models of self-efficacy that can be used at runtime to inform pedagogical decisions. In an empirical study, two families of self-efficacy models were induced: a static model, learned solely from pre-test (non-intrusively collected) data, and a dynamic model, learned from both pre-test data as well as runtime physiological data collected with a biofeedback apparatus. The resulting static model is able to predict students' real-time levels of self-efficacy with reasonable accuracy, while the physiologically informed dynamic model is even more accurate.

1 Introduction

Affect has begun to play an increasingly important role in intelligent tutoring systems. Recent years have seen the emergence of work on affective student modeling [8], detecting frustration and stress [7, 21], modeling agents' emotional states [1, 11, 16], devising affectively informed models of social interaction [13, 18, 20], and detecting student motivation [24]. All of this work seeks to increase the fidelity with which affective and motivational processes are modeled in intelligent tutoring systems in an effort to increase the effectiveness of tutorial interactions and, ultimately, learning.

Self-efficacy is an affective construct that has been found to be a highly accurate predictor of students' motivational state and their learning effectiveness [25]. Defined as "the belief in one's capabilities to organize and execute the courses of action required to manage prospective situations" [2], self-efficacy has been repeatedly demonstrated to directly influence students' affective, cognitive, and motivational processes [3]. Self-efficacy holds much promise for intelligent tutoring systems (ITSs). Foundational work has begun on using models of self-efficacy for tutorial action selection [6] and investigating the impact of pedagogical agents on students' self-efficacy [5, 14]. Self-efficacy is useful for predicting what problems and sub-problems a student will select to solve, how long a student will persist on a problem, how much overall effort they will expend, as well as motivational traits such as level of engagement [22, 25]. Thus, if an ITS could increase a student's self-efficacy, then it could enable the student to be more actively involved in learning, expend more effort, and be more persistent; it could also enable them to successfully cope in situations where they experience learning impasses [3].

To effectively reason about a student's self-efficacy, ITSs need to accurately model self-efficacy. Self-efficacy diagnosis should satisfy three requirements. First, it should be realized in a computational mechanism that operates at runtime. Self-efficacy may vary throughout a learning episode, so pre-learning self-efficacy instruments may or may not be predictive of self-efficacy at specific junctures in a learning episode. Second, self-efficacy diagnosis should be efficient. It should satisfy the real-time demands of interactive learning. Third, self-efficacy diagnosis should avoid interrupting the learning process. A common approach to obtaining information about a student's self-efficacy is directly posing questions to them throughout a learning episode. However, periodic self-reports are disruptive.

This paper reports on the results of an experiment that investigates an inductive approach (naïve Bayes and decision tree classifications) to constructing models of self-efficacy. In the experiment, two families of self-efficacy models were induced: the model learner constructed (1) *static* models, which are based on demographic data and a validated problem-solving self-efficacy instrument [4], and (2) *dynamic* models, which extend static models by also incorporating real-time physiological data. In the experiment, 33 students provided demographic data and were given an online tutorial in the domain of genetics. Next, they were given a validated problem-solving self-efficacy instrument, and they were outfitted with a biofeedback device that measured heart rate and galvanic skin response. Physiological signals were then monitored while students were tested on concepts presented in the tutorial. After solving each problem, students rated their level of confidence in their response with a "self-efficacy slider." Both families of resulting models operate at runtime, are efficient, and do not interrupt the learning process. The static models are able to predict students' real-time levels of self-efficacy with 73% accuracy, and the resulting dynamic models are able to achieve 83% predictive accuracy. Thus, non-intrusive static models can predict self-efficacy with reasonable accuracy, and their predictive power can be increased by further enriching them with physiological data.

The paper is structured as follows. Section 2 discusses the role of self-efficacy in learning. The experimental design is presented in Section 3 (experimental method) and Section 4 (procedure), and the results are described in Section 5. Section 6 discusses the findings and their associated design implications, and Section 7 makes concluding remarks and suggests directions for future work.

2 Self-Efficacy and Learning

Self-efficacy is powerful. It influences students' reasoning, their level of effort, their persistence, and how they feel; it shapes how they make choices, how much resilience they exhibit when confronted with failure, and what level of success they are likely to achieve [2, 22, 25]. While it has not been conclusively demonstrated, many conjecture that given two students of equal abilities, the one with higher self-efficacy is more likely to perform better than the other over time. Self-efficacy is intimately related to motivation, which controls the effort and persistence with which a student approaches a task [15]. Effort and persistence are themselves influenced by the belief the student has that she will be able to achieve a desired outcome [3]. Self-efficacy has been studied in many domains with significant work having been done in computer literacy [9] and mathematics education [19]. It is widely believed that self-efficacy is domain-specific, but whether it crosses domains remains an open question.

A student's self-efficacy[1] is influenced by four types of experiences [3, 25]. First, in enactive experiences, she performs actions and experiences outcomes directly. These are typically considered the most influential category. Second, in vicarious experiences, she models her beliefs based on comparisons with others. These can include peers, tutors, and teachers. Third, in verbal persuasion experiences, she experiences an outcome via a persuader's description. For example, she may be encouraged by the persuader, who may praise the student for performing well or comment on the difficulty of a problem. Her interpretation will be affected by the credibility she ascribes to the persuader. Fourth, with physiological and emotional reactions, she responds affectively to situations. These experiences, which often induce stress and anxiety and are physically manifested in physiological responses, such as increased heart rate and sweaty palms, call for emotional support and motivational feedback.

Self-efficacy holds great promise for ITSs. Self-efficacy beliefs have a stronger correlation with desired behavioral outcomes than many other motivational constructs [10], and it has been recognized in educational settings, that self-efficacy can predict both motivation and learning effectiveness [25]. Thus, if it were possible to enable ITSs to accurately model self-efficacy, they may be able to leverage it to increase students' academic performance. Two recent efforts have explored the role of self-efficacy in ITSs. One introduced techniques for incorporating knowledge of self-efficacy in pedagogical decision making [6]. Using a pre-test instrument and knowledge of problem-solving success and failure, instruction is adapted based on changes in motivational and cognitive factors. The second explored the effects of pedagogical agent design on students' traits, which included self-efficacy [5, 14]. The focus of the experiment reported in this paper is on the automated induction of self-efficacy models for runtime use by ITSs.

3 Method

3.1 Participants and Design

In a formal evaluation, data was gathered from thirty-three subjects in an Institutional Review Board (IRB) of NCSU approved user study. There were 6 female and 27

[1] Self-efficacy is closely related to the popular notion of confidence. To distinguish consider the situation in which a student is very confident that she will fail at a given task. This represents high confidence but low self-efficacy, i.e., she is exhibiting a strong belief in her inability [3].

male participants varying in age, race, and marriage status. Approximately 36% of the participants were Asian, 60% were Caucasian, and 3% were of other races. 27% of the participants were married. Participants average age was 26.15 (SD=5.32).

3.2 Materials and Apparatus

The pre-experiment paper-and-pencil materials for each participant consisted of a demographic survey, tutorial instructions, Bandura's Problem-solving Self-Efficacy Scale [4], and the problem-solving system directions. Post-experiment paper-and-pencil materials consisted of a general survey. The demographic survey collected basic information such as gender, age, ethnicity, and marital status. The tutorial instructions explained to participants the details of the task, such as how to navigate through the tutorial and an explanation of the target domain. Bandura's validated Problem-solving Self-Efficacy Scale [4], which was administered after they completed a tutorial in the domain of genetics, asked participants to rate how certain they were in their ability to successfully complete the upcoming problems (which they had not yet seen). The problem-solving system directions supplied detailed task direction to participants, as well as screenshots highlighting important features of the system display, such as the "self-efficacy slider".

The computerized materials consisted of an online genetics tutorial and an online genetics problem-solving system. The online genetics tutorial consisted of an illustrated 15-page web document which included some animation and whose content was drawn primarily from a middle school biology textbook [17]. The online genetics problem-solving system consisted of 20 questions, which covered material in the online genetics tutorial. The problem-solving system presented each multiple-choice question individually and required participants to rate their confidence, using a "self-efficacy slider," in their answer before proceeding to the next question.

Apparati consisted of a Gateway 7510GX laptop with a 2.4 GHz processor, 1.0 GB of RAM, 15-in. monitor and biofeedback equipment for monitoring blood volume pulse (one sensor on the left middle finger) and galvanic skin response (two sensors on the left first and third fingers). Participants' right hands were free from equipment so they could make effective use of the mouse in problem-solving activities.

4 Procedure

4.1 Participant Procedure

Each participant entered the experimental environment (a conference room) and was seated in front of the laptop computer. First, participants completed the demographic survey at their own rate. Next, participants read over the online genetics tutorial directions before proceeding to the online tutorial. On average, participants took 17.67 (SD = 2.91) minutes to read through the genetics online tutorial. Following the tutorial, participants were asked to complete the Problem-Solving Self-Efficacy Scale considering their experience with the material encountered in the genetics tutorial.

The instrument asked participants to rate their level of confidence in their ability to successfully complete certain percentages of the upcoming problems in the problem-solving system. Participants did not have any additional information about the type of questions or the domain of the questions contained in forthcoming problems. Participants were then outfitted with biofeedback equipment on their left hand while the problem-solving system was loaded. Once the system was loaded, participants entered the calibration period in which they read through the problem-solving system directions. This allowed the system to obtain initial readings on the temporal attributes being monitored, in effect establishing a baseline for HR and GSR.

The problem-solving system presented randomly selected, multiple-choice questions to the participant. The participant selected an answer and then manipulated a self-efficacy slider representing the strength of their belief in their answer being correct. Participants completed 20 questions. Participants averaged 8.15 minutes (SD = 2.37) to complete the problem-solving system. Finally, participants were asked to complete the post-experiment survey at their own rate before concluding the session.

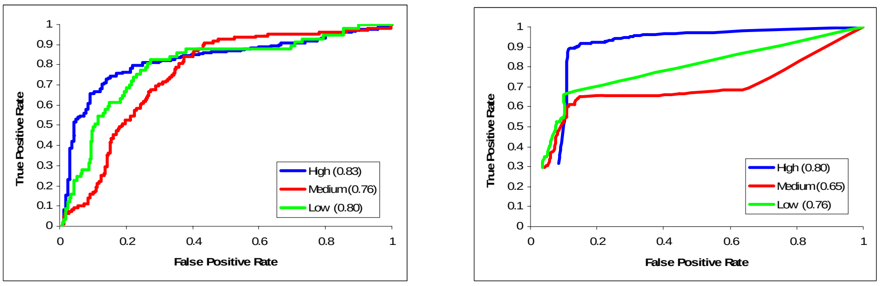
4.2 Machine Learning Procedure

The following procedure was used to induce models of self-efficacy:

- *Data Construction:* Each session log, containing on average 14,645.42 (SD = 4,010.57) observation changes, was first translated into a full observational attribute vector. For example, BVP and GSR readings were taken nearly 30 times every second reflecting changes in both heart rate and skin conductivity.
- *Data Cleansing:* Data were converted into an attribute vector format. Then, a dataset was generated that contained only records in which the biofeedback equipment was able to successfully monitor BVP and GSR throughout the entire training session. Blood volume pulse (used for monitoring HR) readings were difficult to obtain from two participants resulting in the destruction of that data.
- *Naïve Bayes Classifier and Decision Tree Analysis:* The prepared dataset was loaded into the WEKA machine learning package [23], a naïve Bayes classifier and decision tree were learned, and tenfold cross-validation analyses were run on the resulting models. The entire dataset was used to generate several types of self-efficacy models, each predicting self-efficacy with varying degrees of granularity. These included two-level models (Low, High), three-level models, four-level models, and five-level models (Very Low, Low, Medium, High, Very High).

5 Results

Below we present the results of the naïve Bayes and decision tree classification models and provide analyses of the collected data. Various ANOVA statistics are presented for results that are statistically significant. Because the tests reported here were performed on discrete data, we report Chi-square test statistics (χ^2), including both likelihood ratio Chi-square and the Pearson Chi-square values. Fisher's Exact Test is used to find significant p-values at the 95% confidence level ($p < .05$).



(a) (b)

Fig. 1. ROC curves for naïve Bayes (a) and decision tree (b) three-level models of self-efficacy. Areas under the curve are found in the key. Overall the naïve Bayes model correctly classified 72% of the instances while the decision tree was able to correctly classify 83%.

5.1 Model Results

Naïve Bayes and decision tree classifiers are effective machine learning techniques for generating preliminary predictive models. Naïve Bayes classification approaches produce probability tables that can be implemented into runtime systems and used to continually update probabilities for assessing student self-efficacy levels. Decision trees provide interpretable rules that support runtime decision making. The runtime system monitors the condition of the attributes in the rules to determine when conditions are met for assigning particular values of student self-efficacy. Both the naïve Bayes and decision tree machine learning classification techniques are useful for preliminary predictive model induction for large multidimensional data, such as the 144-attribute vector used in this experiment. Because it is unclear precisely which runtime variables are likely to be the most predictive, naïve Bayes and decision tree modeling provide useful analyses that can inform more expressive machine learning techniques (e.g., Bayesian networks) that also leverage domain experts' knowledge.

All models were constructed using a tenfold cross-validation scheme. In this scheme, data is decomposed into ten equal partitions, nine of which are used for training and one used for testing. The equal parts are swapped between training and testing sets until each partition has been used for both training and testing. Tenfold cross-validation is widely used for obtaining a sufficient estimate of error [23].

Cross-validated ROC curves are useful for presenting the performance of classification algorithms for two reasons. First, they represent positive classifications, included in a sample, as a percentage of the total number of positives, against negative classifications as a percentage of the total number of negatives [23]. Second, the area under ROC curves is widely accepted as a generalization of the measure of the probability of correctly classifying an instance [12].

The ROC curves (Fig. 1) above show the results of both a naïve Bayes and decision tree three-level model. Low-confidence was noted by a student self-efficacy rating lower than 33 (on a 0 to 100 scale). Medium-confidence was determined by rating between 33 and 67, while High-confidence was represented all ratings greater than 67. The smoothness of the curve in Figure 1(a) indicates that sufficient data seems to have been used for inducing naïve Bayes models. The jaggedness of the

Table 1. Model results – area under ROC curves. Gray rows indicated static models induced from non-intrusive demographic and Problem-Solving Self-Efficacy data. The other rows represent dynamic models that were also based on physiological data.

Model	Two-level	Three-level	Four-level	Five-level
Naïve Bayes	0.85	0.72	0.75	0.64
Decision Tree	0.87	0.83	0.79	0.75
Naïve Bayes	0.82	0.70	0.69	0.63
Decision Tree	0.83	0.73	0.69	0.64

curves in Figure 1(b) indicates that more data covering the possible instances is needed. In particular, further investigation will need to consider more opportunities for students to experience instances of low self-efficacy. Despite the appearance of a lack of sufficient data, the decision tree model performed significantly better than the naïve Bayes model (likelihood ratio, $\chi^2 = 21.64$, Pearson, $\chi^2 = 21.47$, $p < .05$). The highest performing model induced from all data was the two-level decision-tree based dynamic model, which performed significantly better than the highest performing static model, which was a two-level decision tree model (likelihood ratio, $\chi^2 = 3.99$, Pearson, $\chi^2 = 3.97$, $p < .05$). The three-level dynamic decision tree model was also significantly better than the static three-level decision tree (likelihood ratio, $\chi^2 = 18.26$, Pearson, $\chi^2 = 18.13$, $p < .05$). All model results are presented in Table 1.

5.2 Model Attribute Effects on Self-Efficacy

Heart rate and galvanic skin response had significant effects on self-efficacy predictions (Table 2). Participants' age group was the only demographic attribute to have a significant effect on all levels of self-efficacy models (Table 3).

Table 2. Chi-squared values representing the significance of physiological signals on varying levels of dynamic self-efficacy models ($p < 0.5$). Grayed cells represent no significance.

Physiological signal	Two-level	Three-level	Four-level	Five-level
HR		9.58	15.35	12.78
GSR		9.24	17.96	14.82

Table 3. Demographic effects on self-efficacy. Chi-square values reported with $p < .05$. Grayed cells represent no statistical significance.

Demographic	Two-level	Three-level	Four-level	Five-level
Gender			18.10	11.14
Age Group	16.25	50.00	94.64	87.64
Race & Ethnicity				

6 Discussion and Future Work

Self-efficacy is closely associated with motivational and affective constructs that both influence (and are influenced by) a student's physiological state. It is therefore not unexpected that a student's physiological state can be used to more accurately predict

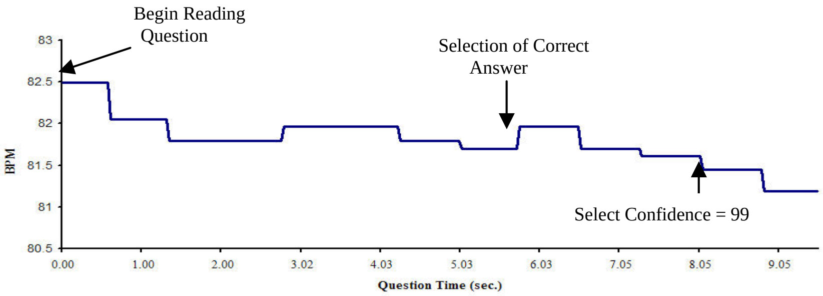


Fig. 2. Heart rate for reported high self-efficacy student.

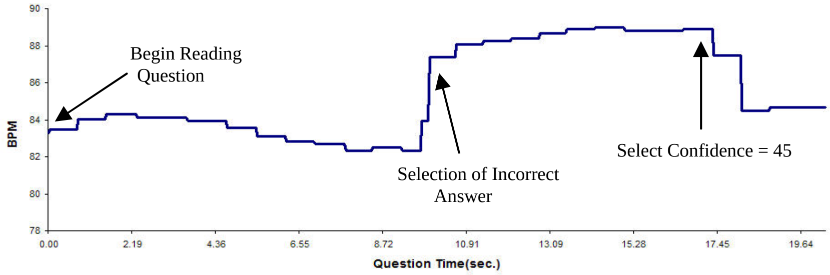


Fig. 3. Heart rate for reported low self-efficacy student.

her self-efficacy. For example, Figures 2 and 3 show the heart rates for one participant in the study over the course of solving two problems. In Figure 2, the participant reported high levels of self-efficacy, while the same participant whose heart rate progression is shown in Figure 3 reported low levels of self-efficacy. The heart rate for the high self-efficacy student gradually drops as they encounter a new question, presumably because of their confidence in their ability to successfully solve the problem. In contrast, the heart rate for the low self-efficacy student spikes dramatically when the student selects an incorrect answer. This phenomenon is particularly intriguing since the students were in fact not given feedback about whether or not their responses were correct. It appears that through some combination of cognitive and affective processes the student's uneasiness with her response, even in the absence of direct feedback, was enough to bring about a significant physiologically manifested reaction. Curiously, there is a subsequent drop in heart rate after the student reports her low level of self-efficacy. In this instance, it seems that providing an opportunity to acknowledge a lack of ability and knowledge to perform may itself reduce anxiety.

The experiment has two important implications for the design of runtime self-efficacy modeling. First, even without access to physiological data, induced decision-tree models can make reasonably accurate predictions about students' self-efficacy. Sometimes physiological data is unavailable or it would be too intrusive to obtain the data. In these situations, decision-tree models that learn from demographic data and data gathered with a validated self-efficacy instrument administered prior to problem solving and learning episodes, can accurately model self-efficacy. Second, if runtime physiological data is available, it can significantly enhance self-efficacy modeling. Given access to HR and GSR, self-efficacy can be predicted more accurately.

7 Conclusion

Self-efficacy is an affective construct that may be useful for increasing the effectiveness of tutorial decision making by ITSs. It may be helpful for increasing students' level of effort, the degree of persistence with which they approach problem solving, and, ultimately, the levels of success they achieve. However, to provide accurate and useful information, self-efficacy models must be able to operate at runtime, i.e., during problem-solving episodes, they must be efficient, and they must avoid interrupting learning. A promising approach to constructing models of self-efficacy is inducing them rather then manually constructing them. In a controlled experiment, it has been demonstrated that *static* models induced from demographic data, a validated self-efficacy instrument, and information from the tutorial system can accurately predict student's self-efficacy during problem solving. It has also been empirically demonstrated that *dynamic* models enriched with physiological data can even more accurately predict student's self-efficacy during problem solving.

The findings reported here contribute to the growing body of work on affective reasoning for learning environments. They represent a first step towards a comprehensive theory of self-efficacy that can be leveraged to increase motivation and learning effectiveness. Two directions for future work are suggested by the results. First, it is important to pursue studies that investigate techniques for achieving the predictive power of dynamic models but "without the wires." Because of the invasiveness of biofeedback apparatus, it would be desirable to develop self-efficacy models that can be induced from students' actions in learning environments that perhaps can be used to infer physiological responses without actually requiring students in runtime environments to be outfitted with biofeedback sensors. Second, now that self-efficacy can be accurately modeled at runtime, the effect of specific pedagogical actions on students' self-efficacy can be investigated. Thus, it may be possible to quantitatively gauge the influence of competing tutorial strategies on students' self-efficacy, which might further increase learning effectiveness.

References

1. André, E., and Mueller, M. Learning affective behavior. In *Proceedings of the 10th International Conference on Human-Computer Interaction*. Lawrence Erlbaum, Mahwah, NJ, 2003, 512-516.
2. Bandura, A. Exercise of personal and collective efficacy in changing societies. In Bandura, A. (Ed.) *Self-efficacy in changing societies* (pp.1-45). New York, NY: Cambridge University Press, 1995.
3. Bandura, A. *Self-efficacy: The exercise of control*. New York: Freeman. 1997.
4. Bandura, A. *Guide for constructing self-efficacy scales*. Unpublished manuscript. 2005.
5. Baylor, A. and Kim, Y. Pedagogical agent design: The impact of agent realism, gender, ethnicity, and instructional role. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, Springer-Verlag, New York, NY, 2004, 592-403.
6. Beal, C. and Lee, H. Creating a pedagogical model that uses student self reports of motivation and mood to adapt ITS instruction. In *Workshop on Motivation and Affect in Educational Software, in conjunction with the 12th International Conference on Artificial Intelligence in Education*. 2005.

7. Burleson, W. and Picard, R. Affective agents: Sustaining motivation to learn through failure and a state of stuck. In *Workshop of Social and Emotional Intelligence in Learning Environments, in conjunction with the 7th International Conference on Intelligent Tutoring Systems*. 2004.
8. Conati, C., and Mclaren, H. Data-driven refinement of a probabilistic model of user affect. In *Proceedings of the 10th International Conference on User Modeling.* Springer-Verlag, New York, NY, 2005, 40-49.
9. Delcourt, M., and Kinzie, M. Computer technologies in teacher education: the measurement of attitudes and self-efficacy. *Journal of Research and Development in Education*. 27(1):35-41, 1993.
10. Graham, S., and Weiner, B. Principles and theories of motivation. In Berliner, D., and Calfee, R. (Eds.) *Handbook of educational psychology* (pp.63-84). New York, NY: MacMillan Publishing, 1996.
11. Gratch, J., and Marsella, S. A domain-independent framework for modeling emotion. *Journal of Cognitive Systems Research*, 5(4):269-306, 2004.
12. Hanley, J. and McNeil, B. The meaning and use of the area under the Receiver Operating Characteristic (ROC) curve. *Radiology* (143):29-36, 1982.
13. Johnson, L., and Rizzo, P. Politeness in tutoring dialogs: “run the factory, that’s what I’d do”. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems.* Springer-Verlag, New York, NY, 2004, 67-76.
14. Kim, Y. Empathetic virtual peers enhanced learner interest and self-efficacy. In *Workshop on Motivation and Affect in Educational Software, in conjunction with the 12th International Conference on Artificial Intelligence in Education*. 2005.
15. Lepper, M., Woolverton, M., Mumme, D., & Gurtner, J. Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. In Lajoie, S. and Derry, S. (Eds.), *Computers as cognitive tools* (pp. 75-105). Hillsdale, NJ: Erlbaum. 1993.
16. Lester, J., Towns, S., and FitzGerald, P. Achieving affective impact: Visual emotive communication in lifelike pedagogical agents. *The International Journal of Artificial Intelligence in Education*, 10(3-4):278-291, 1999.
17. Padilla, M., Miaoulis, I., and Cyr, M. *Science Explorer: Cells and Heredity*. Teacher’s Edition, Prentice Hall, Upper Saddle River, NJ, 2000.
18. Paiva, A., Dias, J., Sobral, D., Aylett, R., Woods, S., Hall, L., and Zoll, C. Learning by feeling: Evoking empathy with synthetic characters. *Applied Artificial Intelligence*, 19:235-266, 2005.
19. Pajares, F., and Kranzler, J. Self-Efficacy beliefs and general mental ability in mathematical problem solving. *Contemporary Educational Psychology*, 20:426-443, 1995.
20. Porayska-Pomsta, K. and Pain, H. Providing Cognitive and Affective Scaffolding through Teaching Strategies, In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*. Springer-Verlag, New York, NY, 2004, 77-86.
21. Prendinger, H., and Ishizuka, M. The empathic companion: A character-based interface that addresses users’ affective states. *Applied Artificial Intelligence*, 19:267-285, 2005.
22. Schunk, D. and Pajares, F. The development of academic self-efficacy. In Wigfield, A. and Eccles, J. (Eds.), *Development of achievement motivation* (pp. 15-31). San Diego, CA: Academic Press. 2002.
23. Witten, I., and Frank, E. *Data Mining: Practical machine learning tools and techniques.* 2nd Edition, Morgan Kaufman, San Francisco, CA, 2005.
24. de Vicente, A., and Pain, H. Informing the detection of the students’ motivational state: an empirical study. In *Proceedings of the 6th International Conference on Intelligent Tutoring Systems*. Springer-Verlag, New York, NY, 2002, 933-943.
25. Zimmerman, B. Self-efficacy: an essential motive to learn. *Contemporary Educational Psychology* 25:82-91, 2000.