

A Field Experiment with an Animated Intelligent Tutor in Portugal

Study results of a pedagogical experimentation intended to replicate previous research outcomes

Gustavo Santos, Joaquim Jorge
INESC-ID / Department of Computer Science and
Engineering
IST, Technical University of Lisbon
Lisbon, Portugal
{gustavossantos, jorgej}@ist.utl.pt

Francisco José García Peñalvo
Departamento de Informática y Automática
Universidad de Salamanca
Salamanca, Spain
fgarcia@usal.es

Abstract—This paper describes the results of a pedagogical experiment using an Intelligent Tutoring System that also implements an Animated Pedagogical Agent. The study was performed in Portugal, as an attempt to replicate some experiments performed in the USA, and in order to verify the usefulness of the Pedagogical Agent in a different country with distinct culture. After running the experiment and analyzing the data, we have concluded that in our specific case, we did not find any evidence that the pedagogical agent could boost instructional outcomes of students in Portugal. Our data is showing no statistically significant differences between the control, and experimental groups. Additionally, we have also found that the control group actually performed slightly better than the experimental group. As a result, our study is giving some good evidences that a Pedagogical Agent that was developed, tested and successfully used in one country, may not be successful when used in a different country.

Keywords—Computers and Education; Intelligent Tutoring Systems; Animated Pedagogical Agents

I. INTRODUCTION

With the constant development of information and communication technologies, the usage of E-Learning tools tends to increase [1]. Every day more schools, universities and also private companies are adopting educational systems to improve the educational process [2]. Several forms of technology-enhanced learning are currently available on the market, and some successful educational tools for improving learning outcomes are the Intelligent Tutoring Systems (ITSs).

Intelligent Tutors are adaptive educational systems that can provide direct, personalized instruction and feedback to students. These systems have been used in several domains, from physics [3] and middle school math [4], to military applications [5] and programming languages [6]. Many studies have demonstrated that ITSs can increase learning outcomes. Accordingly, some well-known ITSs have been used for many years in real world educational settings, and some have even been commercialized in the market, such as the Carnegie Learning Inc. Cognitive Tutors.

Another contemporary and promising topic in learning technologies is Animated Pedagogical Agents (APAs). APAs are lifelike independent characters that basically habit educational systems to create enriched learning interactions. APAs can express emotional responses to tutorial situations, and they are currently considered a great promise for broadening the tutorial communication between educational systems and students.

Although APAs can vary a lot when it comes to their sophistication and forms of design, they all share the fact of having strong visual presence in interactive educational environments. Moreover, recent research has shown concrete evidences that embodied conversational agents can be beneficial for learning [7]–[9]. Accordingly, it is important to carry on some research studies about the conditions where adding an APA to an ITS can improve even more learning outcomes, and in what conditions this may not be true.

For instance, it is common to have students trying to “game” the system while they work with ITSs [10]. Once tutors provide step-based feedback, it is possible for students to introduce arbitrary inputs on steps to see if they are correct. It is also possible to repetitively ask for ITS hints, without even thinking about the solution. Gaming the system consists of repetitively guessing the correct answer of a step, instead of trying to solve it, or abusing of the help provided by the ITS in search for an answer [11].

Recent research has shown that gaming behaviors lead students to benefit less from the instruction [12]. For instance, the Carnegie Learning Inc. Algebra Cognitive Tutor, which had shown to increase learning outcomes is also related to a high incidence of gaming behaviors. For that reason, an educational intervention characterized by an APA named Scooter was intentionality designed to reduce gaming incidence. In the USA, the Scooter intervention was very successful. It reduced gaming behaviors by almost half, and significantly improved the learning outcomes of gaming students [13].

Essentially, the Scooter intervention is an APA that uses natural language to provide some specific feedback to the

The Portuguese Foundation for Science and Technology (FCT) supported this work through the doctoral grant reference SFRH / BD / 66225 / 2009 and pluriannual INESC-ID research grant project PEst-OE/EEI/LA0021/2011.

apprentices. Furthermore, Scooter can react to student actions by displaying emotions while learners are working with the ITS. Initially, Scooter is cheerful and smiling, but it can also demonstrate other emotions. For example, when students are employing gaming behaviors, like clicking fast through the ITS hints, Scooter can gradually display dissatisfaction. On the other hand, when students are having good instructional results, Scooter praises them and encourages them to keep doing a good job.

The initial goal of this project was to replicate in Portugal the experiments performed with the Scooter agent in the USA, in order to verify its usefulness in a different country with distinct culture. We were originally expecting that likewise the US, the intervention would reduce gaming behaviors by almost half, and significantly improve the learning outcomes of gaming students. Unfortunately, due to some technical problems with the ITS logs, we could not determine who were the gaming students. Consequently, it was impossible to fully replicate the study, since the complete replication requires knowing which students tried to game the system.

Despite not being able to fully replicate the American experiments, we could still analyze some student data and get interesting results. For instance, it was not possible calculate the instructional gain separately for each gaming student, as the original experiment calculated. However, using pre and posttest results, we could still determine the instructional gain for students in two different conditions: (1) Scooter condition and (2) No-Scooter condition. As a hypothesis, we evaluated if the experimental condition (the one with the Scooter agent) could present better leaning outcomes when compared to the control condition (the one with the original Carnegie Learning Inc. Algebra Cognitive Tutor - without the Scooter Agent).

After running the experiment and analyzing the data, we have concluded that despite some previous research has shown that APAs can increase learning outcomes, in our specific case we did not find any evidence that the Scooter agent can boost instructional results in Portugal. Accordingly, there are no statistically significant differences between the control, and experimental groups. Both groups performed statistically the same.

We have to stress that the Scooter agent was not intentionally developed to increase the learning outcomes of every student. It was developed to influence only the gaming students, and have no effect on the other students. However, we find that no difference between experimental conditions is relatively unexpected, since both conditions should naturally have gaming students. Accordingly, because the gaming students are expected to present better instructional gains in the Scooter condition, it is likely that the experimental condition would present at least slightly better educational gains. However, our data shows exactly the opposite. The experimental condition even presented slightly negative educational gains, which are not statistically significant, but still intriguing.

To conclude, our data is indicating that adding an APA to an ITS might not make any improvements when it comes to learning outcomes. Specifically, when the APA was developed to be used in one country, but it is used later on by subjects of

another country. These research results are particularly important because our study is giving good clues that it is incorrect to assume that inserting a specific successful APA on an ITS will increase the learning outcomes of distinct populations. Consequently, further studies should be performed to understand how cultural differences maybe influencing learning outcomes on our specific case.

II. DESCRIPTION OF THE EXPERIMENT

The experiment was performed at the “Escola Secundária Don Dinis” (Liceu D. Dinis) in Santo Tirso, at the district of Porto in Portugal. The school Director and one local teacher were involved in the preparation of the study. The school was able to assign two classes of the Portuguese 8th year of instruction to participate in the project.

In total, there were 46 students (23 in the control group and 23 in the experimental group). The average age was 13 years old (approximately 50% man and approximately 50% woman). The study sessions were conducted in the normal period of classes, without significant changes in the regular schedule of the students. The experimental procedure followed the classic settings of the studies in the field of the Learning Technologies. The experimental sessions were: (1) Pre-test, (2) Educational intervention, (3) Post-test, (4) Data analysis.

The math unit used on this project was the “Interpreting and Creating Scatterplots” unit from the Carnegie Learning Inc. Algebra Cognitive Tutor. This is an important subject in math education, at the middle school level, currently receiving particular emphasis in many European countries, and also the USA. On this topic, students make decisions about variable types and their appropriateness for bar graphs and scatterplots. Students can use a graphing tool to define scales, label axes, analyze and create scatterplots, etc. Fig. 1 shows a snapshot of the control condition of the system.

As we can see, there are four main windows: (1) The scenario window, which presents the wording of the problem; (2) The worksheet window, which shows a table with the information about the problem; (3) The form window, which contains the widgets where students can introduce data; (4) The skill window, which shows the proficiency levels of each skill used in the domain.

Basically, students start reading the problem and then start analyzing the data presented on the table. After that, they start introducing values in the input form and get feedback from the ITS on every step. At any moment, students can click on the question mark button, or go to the help menu to ask for a hint. The experimental condition of the ITS is very similar to the control condition, except for the Scooter APA. Fig. 2 shows the interface of the ITS with the Scooter agent.

Initially, Scooter is cheerful and smiling. However, it can demonstrate different emotions. For instance, if students employ gaming behaviors (like clicking fast through the ITS hints), Scooter becomes progressively upset and warns the students that these behaviors are not beneficial for learning (see the middle of Fig. 3 to visualize Scooter getting progressively upset). On normal situations, Scooter gives regular ITS feedback, which is similar to the feedback given by the system in the control condition.

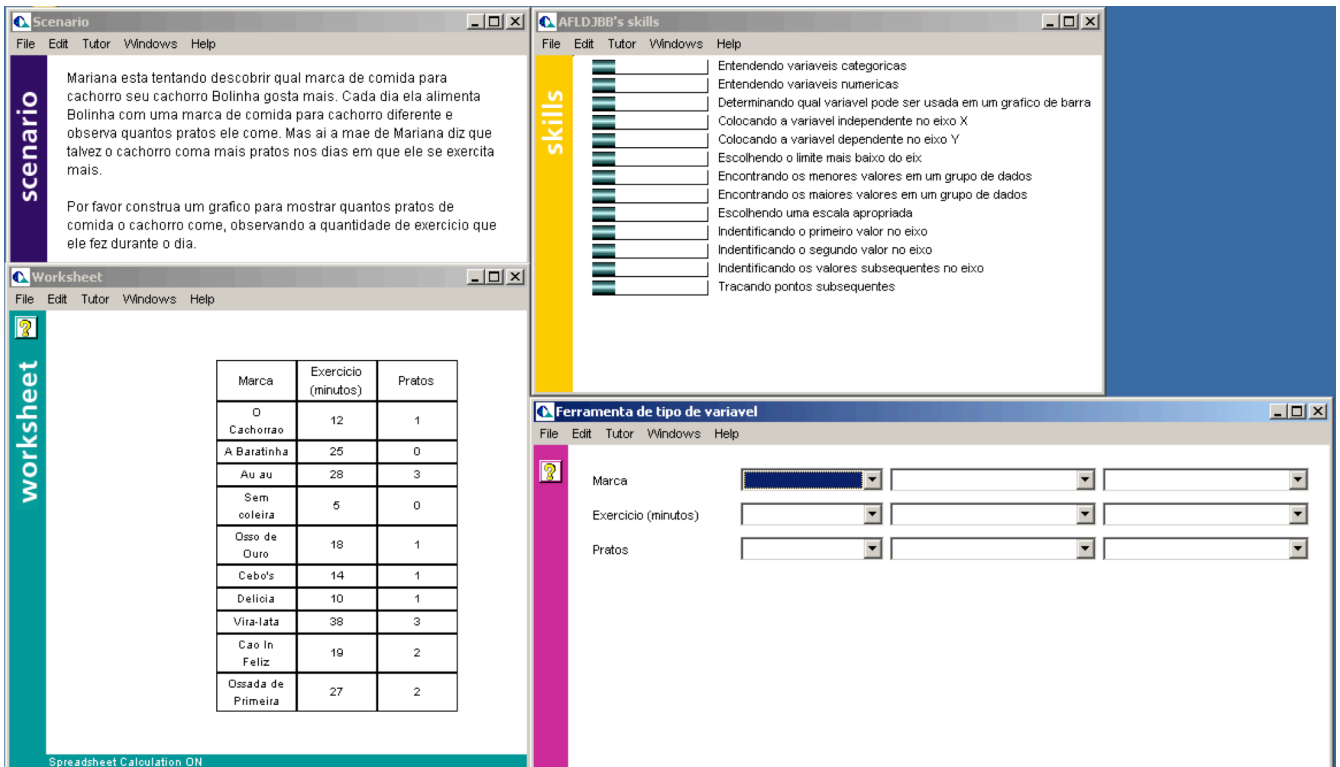


Fig. 1. Snapshot of the ITS Used by the Control Condition.

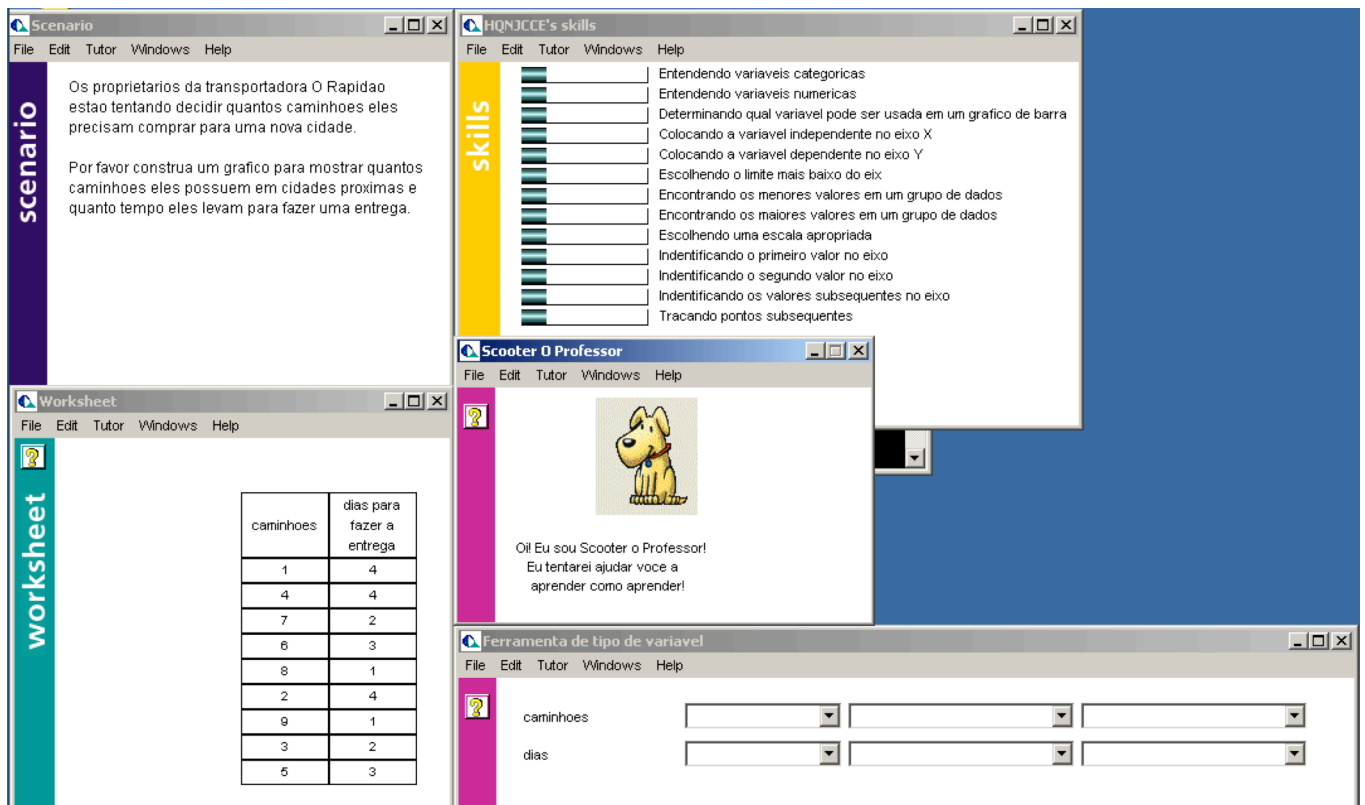


Fig. 2. Snapshot of the ITS Used by the Experimental Condition.

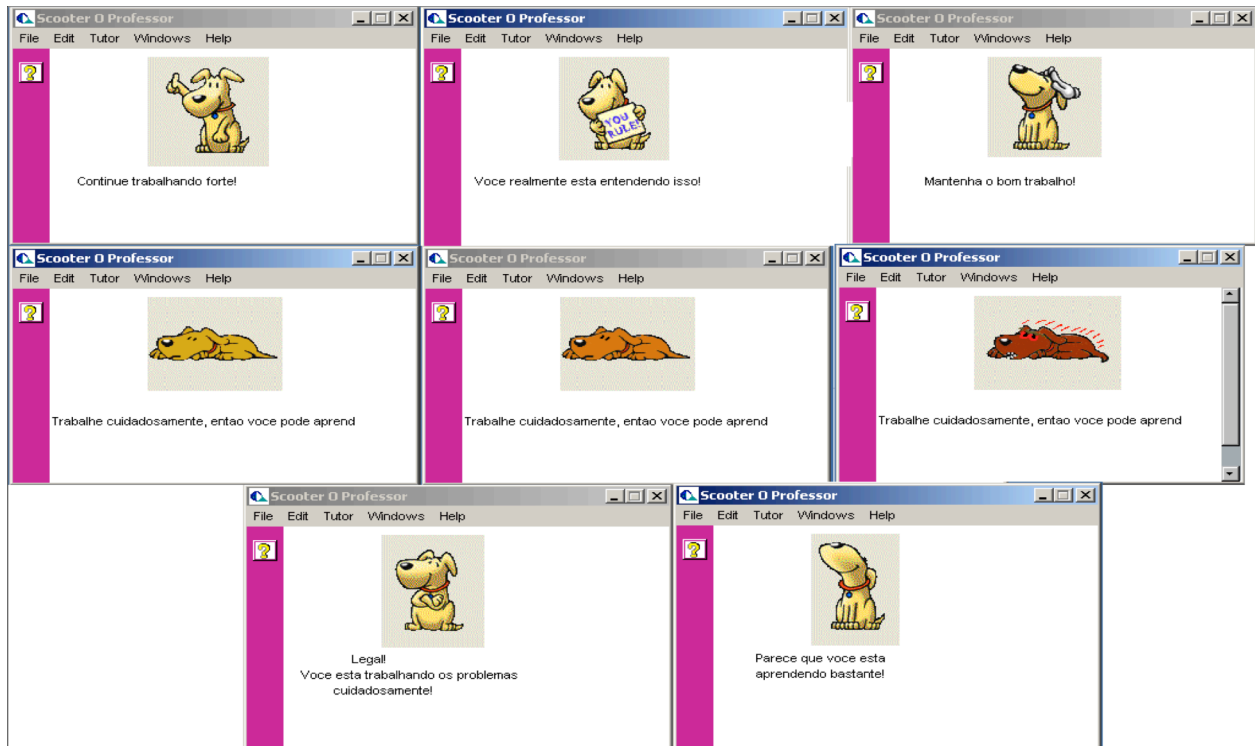


Fig. 3. Some Scooter Reactions. The top three screens depict positive feedback (one is randomly presented for each correct answer). The middle three screens reflect Scooter with increasingly irritated when gaming behaviour is detected. The bottom two screens denote correct responses to remedial exercises presented after gaming was detected.

Scooter's behavior can change when a student is noticed to be gaming the system. A machine-learning detector does the identification of gaming students automatically [12]. If the detector considers that a student has been cheating, Scooter exhibits increasing levels of displeasure, to signal that the student should stop gaming and try to get the answer of the problem in an appropriate way.

These visual expressions of emotion are combined with meta-cognitive messages, proposing that the student should work prudently in order to learn. When a student finds a correct answer through gaming, the APA presents a set of supplementary exercises (see Fig. 4), which were designed to give the student another chance to cover the subject bypassed through gaming.



Fig. 4. Scooter presenting supplementary exercises.

The detector of gaming has been proved to be over 80% correct, when distinguishing gaming and non-gaming students [12]. Inaccuracy is generally evenly distributed between false

negatives and false positives, implying that some students will not receive exercises that would have been appropriate, and some students will receive unnecessary supplementary exercises.

The students that receive unnecessary exercises are not considerably disfavored because they probably know the necessary skills to answer the questions, and can quickly resume working. On the other hand, students that do not know the skills may benefit from the exercises the same way as the gaming students. Therefore, these interventions can be considered "fail-soft," having small penalties for false positives.

In short, Scooter was essentially designed to benefit students in three different ways. First, by indicating how much each student has been gaming, the APA serves as a persistent reminder that students should not game. Second, by expressing emotions when students game the system, the APA displays the natural behavior of dissatisfaction, and it stresses one more time that gaming should not be done. Third, by presenting supplementary exercises, the APA gives the students a second chance to learn some material that they may have entirely missed when applying gaming behaviors.

Finally, it is important to stress that when tested in classrooms in the USA, the Scooter agent have proved to be successful [13]. For instance, students who considerably received supplementary exercises from Scooter had great learning gains, and matched to the rest of the class. In addition, incorporating the APA into the ITS also led to diminish about half the number of students choosing to game.

III. STUDY RESULTS

As mentioned before, the original aim was to replicate in Portugal the experiments performed using the Scooter agent in the USA. Unfortunately, due to technical problems with the ITS logs, we could not determine who were the gaming students. Accordingly, it was not possible to perform the complete replication of the experiment. However, we could still analyze some of the available data, and get particularly interesting results.

First, using pretest and posttest grades, we have determined the instructional gain (posttest score – pretest score) for students in two different conditions: (1) Scooter condition and (2) No-Scooter condition. As a hypothesis, we evaluated if the experimental condition (the one with the Scooter agent) could present better leaning outcomes when compared to the control condition (the one with the original Carnegie Learning Inc. Algebra Cognitive Tutor - without the Scooter Agent).

Educational gain scores were examined in an analysis of variance (experimental vs. control), in order to verify if there were significant differences between the two groups. The increase on scores was greater for participants in the control condition [$M = 0.101449$, $SE = 0.0541829$] than for those in the experimental condition. In fact, there was a very small decrease on scores (negative educational gains) for participants in the experimental condition [$M = -.010870$, $SE = .0464529$]. However, these results are not statistically significant [$F=6.570$, $p=.014$].

Inspection of the 95% confidence intervals around each mean states that there was no significant increase on grades for the participants in the treatment condition and also no significant decrease in grades for the participants in the control condition, as both confidence intervals contain 0. Accordingly, we have to accept the null hypothesis and therefore we cannot infer that there is any statistically significant difference between the two groups, when it comes to educational gains.

Nevertheless, we have to stress that the Scooter agent was not intentionally developed to increase the learning outcomes of every student. It was intent to influence only the gaming students, and have no effect on the other students. However, it is fairly unexpected to have no significant differences between experimental conditions. Additionally, it is even more unexpected that the experimental group presented slightly negative educational gains.

For instance, considering that both conditions should naturally have gaming students, and because the gaming students are expected to present better instructional gains in the Scooter condition, we predicted based on the American experiments that the experimental condition would have at least slightly better educational gains. On the other hand, our data shows exactly the opposite. The experimental condition surprisingly presented slightly negative educational gains, which are not statistically significant, but still intriguing.

However, looking at these results from a different perspective, it is naturally accepted and scientifically proved that different cultures react differently to educational interventions [14]. Accordingly, one intervention that works in one country is not guaranteed to work in a different country. In

fact, our experiment is pointing exactly on this direction, giving us evidences that adding an APA to an ITS might not make any improvements when it comes to learning outcomes, especially when the APA was developed for one country and it is being used by subjects of a different country.

Finally, one important detail about this study that could help to explain the experimental results is that the pedagogical agent was not very appreciated by the students and the math teacher. The teacher once directly stated to the researchers that the Scooter agent could be a little annoying. Additionally, during the tutor sessions, many students have also said the same thing about the APA. Consequently, as participants found the Scooter agent a little irritating, this may have caused the experimental group to present slightly negative educational gains when compared to the control condition.

We believe that North American cultures are more disposed to extroversion and engagement in heavy social interactions, and therefore maybe more tolerant to the poling and the interruptions that are raised by the Scooter agent. On the other hand, Portuguese citizens are more reserved and not so predisposed to disruptions. This may be the reason why the educational intervention was not seemingly successful in Portugal. However, to be precise about why the APA does not seem to work with Portuguese students, further experiments followed by a deeper analysis should be performed to investigate the root causes of these results.

IV. CONCLUSIONS AND FUTURE WORK

This document describes a field experiment that evaluated in Portugal the effects of using an educational intervention that was initially designed for the USA. An APA named Scooter, which is a cartoon puppy that uses natural language to provide instructional feedback, while visually displaying emotions, attempts to discourage gaming behaviors by students. Scooter is thus intended to prevent these behaviors, while increasing the learning outcomes of gaming students. In the USA, this intervention presented good educational results.

While trying to replicate the American experiments in Portugal, we encountered a few problems with the experiment logs and therefore could not fully replicate the original study. However, we could still analyze the available data and conclude that the intervention did not produce any statistically significant differences on learning outcomes as compared to the control condition. Additionally, the experimental condition presented slightly negative educational gains when compared to the control condition. Therefore, we have some good indications that the intervention was not successful in Portugal, maybe due to ethnical and cultural differences regarding roles in the classroom.

To conclude, this research is relevant because many studies in learning technologies tend to generalize local results, as if they would be applicable to different geographical locations, ethnical or cultural environments. However, as we can see from our data, this may not be so easy to accomplish. As future work, it is important to carry on a deeper analysis to determine why this and other similar educational interventions seemingly do not work in different environments. Indeed, further studies are warranted to ascertain what factors might lead to different

results, and how and why these factors may be interfering on learning outcomes for different students. This knowledge might be instrumental in devising intercultural and interoperable approaches to learning environments.

REFERENCES

- [1] García Peñalvo, F. J. (Ed.) "Advances in E-Learning: Experiences and Methodologies". Hershey, PA, USA: Information Science Reference (formerly Idea Group Reference), ISBN 978-1-59904-756-0. January 2008
- [2] B. Beatty and C. Ulasewicz, "Faculty Perspectives on Moving from Blackboard to the Moodle Learning Management System," *TechTrends*, vol. 50, no. 4, pp. 36–45, 2006.
- [3] K. Vanlehn, C. Lynch, K. Schulze, J. Shapiro, R. Shelby, L. Taylor, D. Treacy, A. Weinstein, and M. Wintersgill, "The Andes physics tutoring system: five years of evaluations," in *In Proceedings of the 12th international conference on Artificial Intelligence in Education*, 2005, pp. 678–685.
- [4] S. Ritter, J. R. Anderson, K. R. Koedinger, and A. Corbett, "Cognitive tutor: Applied research in mathematics education," *Psychon. Bull. Rev.*, vol. 14, no. 2, pp. 249–255, 2007.
- [5] J. E. McCarthy, "Military Applications of Adaptive Training Technology," in *Technology Enhanced Learning: Best Practices*, D. G. Lytras P. Ordóñez de Pablos, & W. Huang, Ed. IGI Global, 2008, pp. 304–347.
- [6] A. T. Corbett and J. R. Anderson, "The LISP intelligent tutoring system: Research in skill acquisition," in *Computer Assisted Instruction and Intelligent Tutoring Systems: Establishing Communication and Collaboration*, J. Larkin and R. Chabay, Eds. Hillsdale, 1992.
- [7] J. Lester, S. Converse, B. Stone, S. Kahler, and T. Barlow, "Animated pedagogical agents and problem-solving effectiveness: A large-scale empirical evaluation," in *In Proceedings of Eight World Conference on Artificial Intelligence on Education*, 1997, pp. 23–30.
- [8] R. Lowe and W. Schnotz, *Learning with Animation: Research Implications for Design*. Cambridge University Press, 2008.
- [9] R. Moreno, R. E. Mayer, H. A. Spires, and J. C. Lester, "The Case for Social Agency in Computer-Based Teaching: Do Students Learn More Deeply When They Interact With Animated Pedagogical Agents?," *Cogn. Instr.*, vol. 19, no. 2, pp. 177–213, 2001.
- [10] R. S. Baker, A. T. Corbett, K. R. Koedinger, and A. Z. Wagner, "Off-task behavior in the cognitive tutor classroom: when students 'game the system'," *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, Vienna, Austria, pp. 383–390, 2004.
- [11] R. Baker, J. Walonoski, N. Heffernan, I. Roll, A. Corbett, and K. Koedinger, "Why Students Engage in 'Gaming the System' Behavior in Interactive Learning Environments," *J. Interact. Learn. Res.*, vol. 19, no. 2, pp. 185–224, 2008.
- [12] R. S. Baker, A. T. Corbett, I. Roll, and K. R. Koedinger, "Developing a generalizable detector of when students game the system," *User Model. User-adapt. Interact.*, vol. 18, no. 3, pp. 287–314, 2008.
- [13] R. S. J. d Baker, A. T. Corbett, K. R. Koedinger, S. E. Evenson, I. Roll, A. Z. Wagner, M. Naim, J. Raspat, D. J. Baker and J. Beck, "Adapting to when students game an intelligent tutoring system," *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, 2006, pp. 392 – 401.
- [14] B. Hirshman, A. Birukou, M. Martin, M. Bigrigg, and K. Carley, "The Impact of Educational Interventions on Real & Stylized Cities," 2008.