Exploring the Implications of Tutor Negativity Towards a Synthetic Agent in a Learning-by-Teaching Environment

Ma. Mercedes T. RODRIGO[1], Regina Ira Antonette M. GELI[1], Aaron ONG[1], Gabriel Jose G. VITUG[1], Rex BRINGULA[2], Roselle S. BASA[2], Cecilio DELA CRUZ[2], Noboru MATSUDA[3]
[1]Ateneo Laboratory for the Learning Sciences, Ateneo de Manila University, Philippines
[2]University of the East, Manila, Philippines
[3]Carnegie Mellon University, Pittsburgh, PA, USA
mrodrigo@ateneo.edu, riamgeli@gmail.com, icemanfresh@yahoo.com, gjgv124@gmail.com, rexbringula@gmail.com, roselle_basa@yahoo.com, ceciledc22@yahoo.com, noboru.matsuda@cs.cmu.edu

ABSTRACT

We examine the implications of negativity in free-form dialogue between student tutors and a synthetic agent in APLUS, a learning-by-teaching online learning environment for Algebra. We attempt to determine whether the negativity of a student tutor's discourse with the agent indicates that the student is learning more or less of the material and whether the feedback they give the synthetic agent is more or less accurate. We found a weak negative correlation between tutor negativity and learning gains and a strong negative correlation between tutor negativity and accuracy of feedback. Negativity might indeed indicate that student tutors lack mastery of the subject matter and need assistance themselves and detecting negativity during tutoring and providing appropriate assistance might enhance the effectiveness of APLUS and other intelligent tutoring systems.

I. INTRODUCTION

In recent years, learning scientists and researchers have found evidence that the nature of the spoken or written discourse that transpires between tutors and tutees has an effect on learning outcomes. In human-to-human tutoring, expert human tutors support their students by showing empathy and warmth. They avoid overt criticism, choosing instead to express confidence in their students' abilities to succeed [1]. This does not mean that discourse has to stay uniformly positive throughout the tutoring relationship. Tickle-Degnen and Rosenthal [2] found that positivity is more crucial in early interactions while the quality of later interactions weighs more heavily on coordination and attentiveness.

These same observations characterize peer-to-peer tutoring. A study showed that when tutor and tutee were friends, learning was increased when they were negative or impolite with each other [3]. Interpersonal conflict—expressed as insults, condescensions, dismissals, curses, and criticisms—has been shown to co-occur with positive cognitive conflict, can increase closeness, and lead to greater learning [4, 5].

Discourse on positivity or negativity continues to have an effect in tutoring contexts between human students and synthetic pedagogical agents. [6] and [7] found that students posted higher learning gains and greater self-efficacy when they received polite feedback, rather than direct feedback only, from a synthetic tutor. Polite feedback was particularly important with low ability or high extroversion students [7]. In learning-by-teaching scenarios [8-9] where the human student acts as the tutor to a less-able synthetic pedagogical agent, the use of playful face-threatening comments and teasing correlate with tutor learning [10]. Indeed, the tutor was less likely to learn if his/her tutoring dialog was highly formal tutoring dialog, implying disconnection between the tutor and tutee.

Not all impolite or negative feedback is constructive. Excessive or overly harsh criticism sabotages both social and cognitive goals [3, 4]. Furthermore, rudeness as a teaching strategy is only effective among friends. Face-threatening discourse between strangers in a learning-by-teaching peer tutoring environment is negatively correlated with learning for both the tutor and the tutee [3].

In this study, we turn our attention to a gap we perceived in the literature: the implications of tutor negativity regarding the tutor learning and tutoring quality in learning-by-teaching situations. When a tutor is harshly critical of or rude to a tutee, is the tutor learning more or less? Is the content that he or she communicates to the learner more or less accurate? Prior work has already shown that some affective states have compromise learning, e.g. boredom or confusion precede or co-occur with non-learning behaviors [11]. We hypothesize that tutor negativity is an indication that the tutor is having difficulty with the subject matter or is frustrated with the tutee's behavior. To test this hypothesis, we record and analyze student self-explanations as they use SimStudent, a learning-by-teaching environment for algebra. We classified their self-explanations first in terms of content then in terms of valence of the affective state exhibited. We then correlated the proportions of self-explanations in the various classifications

We thank the Department of Science and Technology Philippine Council for Industry, Energy, and Emerging Technology Research and Development (PCIEERD) for the grant entitled, "Development of Affect-Sensitive Interfaces," the Engineering Research and Development for Technology (ERDT) program for the grant entitled, "Development of an Educational Data Mining Workbench," and the University of the East.

against student post-test gains, delayed post-test gains, and percentage of steps correctly tutored.

II. Online Learning Environment with SimStudent

SimStudent is a teachable agent that helps students learn linear equation problem-solving skills by teaching [12]. It has been tested and redesigned several times, resulting in insights regarding the effects of learning by teaching and related cognitive theories to explain when and how students learn by teaching—including the effect of self-explanation for tutor learning [13], motivational factors within the competitive game show [14], the effect of formative evaluation for learning by teaching, and some other cognitive and social factors that correlate with tutor learning [15].

This section provides a brief overview of SimStudent and the online learning environment, Artificial Peer Learning environment using SimStudent (APLUS), in which students learn to solve algebra equations by interactively teach SimStudent. Technical details about SimStudent and APLUS can be found elsewhere [15].

A. SimStudent

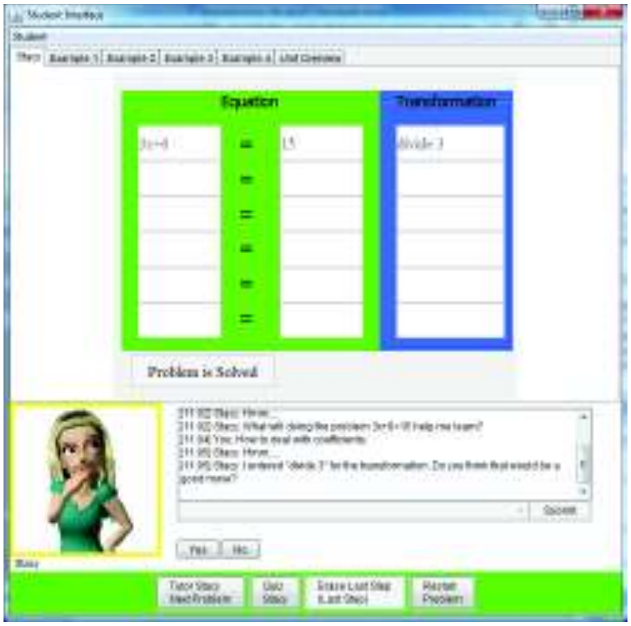
SimStudent is a synthetic pedagogical agent that acts as a peer learner. It learns procedural skills from examples. That is, a student gives SimStudent a problem to solve. SimStudent then attempts to solve the problem one step at a time, occasionally asking the human tutor about the correctness of each step. If SimStudent cannot perform a step correctly, it asks the student for a hint. To respond to this request, the student has to demonstrate the step.

Students are not always able to provide the correct feedback and hints. As SimStudent is unable to distinguish correct from incorrect feedback, it continues to try to generalize examples, generating production rules that represent the skills learned. SimStudent is also capable of making incorrect induction that would allow SimStudent to learn incorrect productions. This is one of SimStudent's unique characteristics: its ability to model students' incorrect learning

B. APLUS: Artificial Peer Learning Environment using SimStudent

In APLUS, students act as a tutor to SimStudent, visualized at the lower left corner of the screen and named Stacy (Figure 1). The tutoring interface allows the student and Stacy to solve problems collaboratively. In the figure, a student poses the problem 3x+6=15 for Stacy to solve. Stacy enters "divide 3" and asks the student whether this is correct. The student responds by clicking on the [Yes/No] button. If the student gets stuck, she can consult the examples tabbed at the top of the screen.

The student has the option of gauging how much Stacy has learned with the use of a quiz. The student chooses when and how often to administer the quiz by clicking a button at the bottom of the interface. The quiz interface looks like the tutoring interface, however, when Stacy takes the quiz, she does so independently, without any feedback or intervention from the student. At the end of the quiz, the student is presented with a quiz result. The quiz is divided into four (4) sections, each with two equation problems. Stacy cannot progress to a section until she passes the previous section. The students were asked to tutor Stacy to be able to solve equations with variables on both sides. In the classroom studies, the students were informed that their goal was to help Stacy pass all four (4) sections of the quiz.



[illegible] Stacy: Hmm...
[illegible] Stacy: What will doing the problem 3x+6=15 help me learn?
[illegible] You: How to deal with coefficients
[illegible] Stacy: Hmm...
[illegible] Stacy: I entered "divide 3" for the transformation. Do you think that would be a good move?

Fig 1. A screen shot of APLUS. SimStudent is visualzed with an avatar image and names Stacy.

C. Self-explanation in APLUS

Two versions of SimStudent were used for this study: an experimental condition in which Stacy prompted the participants to self-explain their tutoring decisions and a control condition with no self-explanation prompts. In the self-explanation condition, Stacy would ask questions like, "Why should I do this problem?" or "But I tried that move earlier. Why doesn't it work now?" Students could then choose a response from a drop-down list or create their own freeform explanations.

The students were not told that the version of Stacy used did not actually understand these explanations nor use them as inputs to the agent's learning process (although some students might have noticed the limitation of SimStuent). Stacy did not process the self-explanations beyond simply recording them. It is this self-explanation that student entered that we analyzed to understand how the negativity of students' response affect tutor learning.

III. Methods

A. Participants

The study took place in one high school in Manila, Philippines, under the supervision of the co-authors from the University of the East and the Ateneo de Manila University.

We enlisted participation from five (5) first year high school sections with an average of 40 students per class. All students were taking an algebra class. There were 187 study participants in all with ages ranging from 11 to 15. The average age of the participants was 12.5 years.

B. *Structure of the study*

Each student was assigned to one of two versions of SimStudent (baseline and self-explanation, as described in II.C). For three consecutive days, participants used their assigned version of SimStudent for one classroom period (60 minutes) per day.

C. *Measures*

The system automatically logged all of the participants' activities including problems tutored, feedback provided, steps performed, examples reviewed, hints requested, and quiz attempts.

As mentioned earlier, Stacy sometimes asks the student whether certain actions are correct. The student responds by clicking a Yes or No button. When the student response is recorded, the system also marks the correctness of the feedback as correct, i.e. the student said Yes when Stacy was indeed correct or No when Stacy was wrong, or incorrect, i.e. the student said Yes when Stacy was wrong or No when Stacy was correct. To arrive at the percentage of steps that the student tutored correctly was computed as the total number of correct responses for the student divided by the total number of responses to Stacy's inquiries of this nature.

Students took pre- and post-test before and after the intervention. The students also took a delayed post test two weeks after the post-test was administered. Three versions of isomorphic tests, tests A, B, and C, were used to counterbalance the pre-, post-, and delayed post tests. As shown in Table I, all three tests had good reliability scores. Cronbach's alpha scores for Test A=0.92 and 0.92; Test B=0.91 and 0.94; C=0.95 and 0.95 as pre-tests and post-tests respectively. 146 out of 187 participants took all three tests.

Each test consists of five parts: (1) six equation solving items where students were asked to show their work on a piece of paper. (2) 38 TRUE or FALSE questions where students had to identify the constant and variable terms in an expression and indicate whether two given expressions are like terms. (3) 12 AGREE or DISAGREE questions in which students had to identify if a given operation is appropriate for a given equation. (4) 10 YES or NO questions in which students had to identify whether a pair of expressions was equivalent. (5) Five items which a mixture of multiple choice and free response questions in which students had to identify and explain an incorrect step for a given equation.

Parts 1, 3, and 5 constituted procedural knowledge while parts 2 and 4 constituted conceptual knowledge. Because there was no main effect of the test-time (pre, post, delayed) for the conceptual knowledge test, we only use the procedural test scores as the learning outcome measure for the current analysis.

In scoring the test, we gave each correct answer 1 point and each wrong answer 0 points. We computed the post-test and delayed post-test normalized gains using the formula:

$$\text{Normalized gain} \qquad \frac{\text{Post-test score} - \text{Pre-test score}}{1 - \text{Pre-test score}} \qquad (1)$$

There were 78 students who have both complete test scores and log data in the self-explanation condition, and those are the students included in the analyses that follow.

IV. Test and Quiz Results

The mean scores of the procedural skill test and their standard deviations are shown in Table II. The results of the pre-test showed that the students in the study had relatively weak prior knowledge. They did post learning gains from the pre-test to the post-test and from the pre-test to the delayed post-test. However, only the gains from the pre-test to the delayed post-test were significant (t(77)=-3.52, p<.001). Students took classroom instructions for two weeks between post-test and delayed-test, which arguably explains the increase of the test score from pre to delayed.

TABLE I. Mean Test Scores ± SD for pre, post, delayed procedural skill test

Pre-test	*Post-test*	*Delayed Post-test*
0.21±0.01	0.23±0.02	0.28±0.03

Students could quiz Stacy as many times as they liked, however, only 29 of the SimStudents were able pass even the first part of the quiz. None of the SimStudents were able to pass the second part of the quiz.

V. Data labeling and Analysis, Round 1

The 78 students in the self-explanation group generated a total of 139,982 transactions (including, for example, quiz attempts, feedback provided, problems tutored, and so on), of which 1,810 were self explanations.

We first attempted to determine if there was a relationship between the quality of the content of the self-explanations and tutor achievement and accuracy. Working independently, three of the authors labeled each self-explanation using the coding schema used in the past study [13] as shown in Table III.

TABLE II. Coding Conventions

Code	Description
N1	The input must include a math concept-oriented explanation or justification of why the student entered the problem , why SimStudent's performance was wrong, or why the student did a particular demonstration. It may also contain a math-related explanation of how to solve the problem, e.g. SimStudent: But before I did 4 for the result of 16 and 4. I thought that would be the same with 3x here. Why is this different? Student input: Because you divide 3 from x and 6 the x is isolated and then you got x=2

Code	Description
N2	The input only provides a math-related explanation of how to solve the problem, e.g. SimStudent question: What is it about a/4=4 that made you know to divide 4? Student input: Multiply 4 by 4 divide 16 by 4 and get your answer
N3	The input blames SimStudent for an incorrect action on the current problem solving process. It does not include a math-related explanation of how to solve the problem or a math-concept-oriented justification for the student's action, e.g. Student input: You didn't add it right.
N4	The input is related to math but is vague and abstract. It does not include a math-concept-oriented justification for the student's action. SimStudent's question: What will doing the problem 6x-2=16 help me learn? Student input: to do these types of equations
N5	The input is an admission on the part of the student that he/she made a mistake, e.g. Student input: Because I was wrong.
N6	The input is an admission on the part of the student that he/she does not know the answer to SimStudent's question Student input: I don't know.
N7	The input does not address SimStudent's question or the input is just a number. SimStudent's question: But before I did 3 for the result of divide 3 and 9. I thought that would be the same with 2 here. Why is this different? Student's input: How to deal with more terms
N8	The input does not fit into the other categories. SimStudent's question: Why did you choose 7y=49 for the problem? Student's input: because I did.

The final label assigned to each self-explanation was label that the majority of coders assigned to that case. When all three coders gave different labels to a self-explanation, the coders convened, discussed, and arrived at a consensus as to what a self-explanation's labels should be. For each student, we computed the percentage of self-explanations for each of the categories N1 through N8. We then correlated these percentages with post-test gains, delayed post-test gains, and percentage of steps that the student tutored correctly. Table IV shows the results of the correlations.

TABLE III. CORRELATIONS BETWEEN SELF-EXPLANATION CATEGORIES, POST-TEST GAINS, DELAYED POST-TEST GAINS, AND PERCENTAGE OF STEPS CORRECTLY TUTORED. P-VALUES ARE IN PARENTHESES

Category	*Post-test gain*	*Delayed Post-Test Gain*	*Percentage of Steps Correctly Tutored*
N1	0.01 (.96)	0.11 (.33)	0.13 (.24)
N2	0.06 (.61)	0.20(.07)	0.21 (.06)
N3	-0.05 (.65)	0.04 (.73)	0.15 (.18)
N4	-0.04 (.74)	0.00 (.99)	0.12 (.30)
N5	-0.02 (.84)	0.03 (.80)	0.08 (.49)
N6	0.03 (.77)	0.15 (.19)	-0.01 (.95)
N7	0.11 (.33)	0.02 (.84)	-0.02 (.86)
N8	-0.09 (.46)	-0.23 (.04)	-0.27 (.02)

The incidence of N2 labels had weak positive correlations with delayed post-test gains (r=.20, p=.07) and percentage of steps correctly tutored (r=.21, p=0.06). This implies that students who give some feedback with math content are more likely to attain higher gain from pre to delayed-test on the procedural skill test. Their feedback to SimStudent also tends to be more accurate.

The incidence of N8 labels, on the other hand had also small negative correlations with delayed post-test gains (r=-.23, p=.04) and percentage of steps correctly tutored (r=-.27, p=.02). This implies that students who give unrelated feedback are more likely to learn less and give poorer quality feedback.

The weakness of the correlations prompted us to relabel the self-explanations with fewer categories. Instead of focusing on content, we chose to focus on the affective state that the student expressed in his or her feedback to SimStudent.

VI. DATA LABELING AND ANALYSIS, ROUND 2

Working independently two of the authors of this paper coded each self-explanation as "negative" or "non-negative" (i.e. positive or neutral self-explanations) . We considered a self-explanation as negative if the student expressed inadequacy in their own knowledge, anger or frustration, by "shouting" at Stacy by typing in all caps (ADD 3B), using insulting or demeaning language (*you know what!? your* [sic] *the worst student I've ever taught in my whole life!!!!!!!!!!!!!!!!!!!!!!!!!!!!*), refusing to offer Stacy assistance (recorded in the logs as "Explanation Not Given"), admitting a lack of knowledge (*I don't know thw* [sic] *answer sorry :(*), or giving unrelated or irrelevant responses (*w3w* – this is Filipino online gamer-speak for "wow").

We considered a self-explanation to be non-negative if it was neutral (*add 2; yes; no*), polite (*ok i help you learn*), or helpful without being irritable (*because you still need to multiply both sides by 5*)

The two coders coded each self-explanation and had an inter-rater reliability of 0.68 [16].

There was an almost even split between the negative and non-negative self-explanations. Of the 3620 codes assigned (1810 per coder), 1809 were negative, while 1811 were non-negative.

For each student, we counted the number of self-explanations for which both coders coded as "negative." We then computed each student's negative self-explanation score (NSE) as the ratio of negative self-explanations to the total number of self-explanations made.

We correlated each student's NSE against his/her post-test gains, delayed post-test gain, and percentage of steps that the student tutored correctly.

We found a small negative correlation between average NSE and delayed post-test gains (r=-.238; p=.037). We also found a medium negative correlation between average NSE and percentages of correctly tutored problems (r=-.424, p<.001). Average NSE were not correlated with post-test gains (r=-.065; p=.574). Tutors who express anger or frustration towards SimStudent tend to give be less accurate and tend to attain less gain from pre to delayed-test on the procedural skill test than those who express less negativity.

Surprisingly enough, we found no significant relationship between prior knowledge, as measured by the pretest, and negativity (r=-1.08; p=.35). This means that what students already knew before using Stacy had no detectible impact on how they related with the her.

VII. Discussion

Much has already been written regarding the role and effects of politeness, positivity, and negativity in human-to-human or human-to-computer tutoring dialogues. Prior literature has shown that both politeness and well-placed rudeness can have positive effects on learning for both tutors and tutees. This paper explores an aspect of negative tutor feedback that is less discussed in the literature: What rudeness implies about the tutor's learning and the tutor's correctness.

We first labeled the self-explanations based on their math-related content. We found that students who give SimStudent content-related help were more likely to do well in the delayed post-test. They also tended to be more accurate in their coaching overall, though both relationships were weak. On the other hand, students who tended to give irrelevant answers were more likely to do poorly in the delayed post-test and give less accurate answers.

When we relabeled the data based on the negativity or non-negativity that the students exhibited in their feedback, we found a weak negative correlation between the percentage of negative comments made by a student tutor and the student tutor's learning gains. Of greater interest was that we found a strong negative correlation between a student tutor's percentage of negative comments and the correctness of the student tutor's feedback.

The implication is that the students were struggling with the subject matter. The low pre-test scores are testimony to their lack of familiarity with linear equations. Upon further discussion with the math teachers, we found that, although the students used SimStudent in the fifth month of their Algebra class, they had still not taken linear equations. Indeed, they only learned how to solve linear equations prior to the delayed post-test. This may have accounted for the general improvement in the delayed post-test scores, but it does not account for the accuracy of students' feedback during their interactions with SimStudent. It is possible that, because of their lack of prior knowledge, the students themselves did not feel confident about what they were teaching, leading to frustration with and the hostility towards Stacy. These findings are consistent with currently ongoing research on student engagement in reading. There are findings that suggest that students disengage with the learning task when the reading material is either too difficult or too easy (Arthur Graesser, *Personal communications*).

The learning-by-teaching paradigm is reputed to be effective because it forces human tutors to gain deeper understanding of the material, to structure and organize the material, and to identify what parts of the material are most important (see [8, 9]). What these findings imply is that students need to have a minimum level of competence in the subject matter before they can assume the role of tutor. The absence of this competence leads to negativity.

Whether we can use tutor negativity as an indicator of lack of mastery is an area that may warrant further research. Learning systems that allow free-form dialog inputs may be able to use this as one feature among others to detect when students are feeling frustrated or angry and disengaging with the subject matter.

Furthermore, the fact that Stacy could not actually respond to the feedback did nothing to mitigate these negative feelings. She responded to all student feedback with the same stoicism, regardless of positivity or negativity. Subsequent versions of Stacy and other similar learning-by-teaching synthetic agents may need to respond to overly negative feedback, to bring the student back to a constructive learning dialog.

Acknowledgment

The authors thank Ryan S. Baker for his helpful comments and suggestions. We thank the Ateneo de Manila University, the Department of Information Systems and Computer Science, Marc Lester Armenta, John Paul Contillo, Victoria Keiser, Jessica O. Sugay, and Evelyn Yarzebinski.

References

[1] M.R. Lepper and M. Woolverton, "The wisdom of practice: Lessons learned from the study of highly effective tutors," in Inspiring AcademicAchievement, J. Aronson (Ed.), pp. 135-159, 2002.

[2] L. Tickle-Degnen, and R. Rosenthal, The nature of rapport and its nonverbal correlates. *Psychological Inquiry*, *1*(4), 285-293, 1990.

[3] A. Ogan, S. Finkelstein, E. Walker, R. Carlson, and J. Cassell, "Rudeness and rapport: Insults and learning gains in peer tutoring," in *Intelligent Tutoring Systems Lecture Notes in Computer Science*, vol. 7315, pp. 11-21, 2012.

[4] D. N. Prata, R. S. J. d. Baker, E. Costa, C. P. Rose, Y. Cui, A. M. J. B. de Carvalho, "Detecting and Understanding the Impact of Cognitive and Interpersonal Conflict in Computer Supported Collaborative Learning Environments. *Proceedings of the 2nd International Conference on Educational Data Mining*, pp. 131-140, 2009.

[5] W. Y. Wang, S. Finkelstein, A. Ogan, A. W. Black, and J. Cassell, "'Love ya, jerkface': Using sparse log-linear models to build positive (and impolite) relationships with teens," *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 2012.

[6] N. Wang and W. Lewis Johnson, "The Politeness Effect in an intelligent foreign language tutoring system," in *Intelligent Tutoring Systems*, pp. 270-280, Springer Berlin Heidelberg, 2008.

[7] N. Wang, W. L. Johnson, R. E. Mayer, P. Rizzo, E. Shaw, and H. Collins, "The politeness effect: Pedagogical agents and learning gains," *Frontiers in Artificial Intelligence and Applications* 125, pp. 686-693, 2005

[8] D. B. Chin, I. M. Dohmen, B. H. Cheng, M. A. Oppezzo, C. C. Chase, and D. L. Schwartz, "Preparing students for future learning with Teachable Agents," *Educational Technology Research and Development* 58, no. 6, pp. 649-669, 2010.

[9] J. A. R. Uresti and B. du Boulay, "Expertise, motivation and teaching in learning companion systems," *International Journal of Artificial Intelligence in Education* 14, no. 2 pp. 193-231, 2004.

[10] A. Ogan, S. Finkelstein, E. Mayfield, C. D'Adamo, N. Matsuda, and J. Cassell, "Oh dear Stacy!: Social interaction, elaboration, and learning with teachable agents," in *Proceedings of the 2012 ACM annual*

conference on Human Factors in Computing Systems, pp. 39-48. ACM, 2012.

[11] M. M T. Rodrigo, R. S. J. d. Baker, M. C. V. Lagud, S. A. L. Lim, A. F. Macapanpan, S. A. M. S. Pascua, J. Q. Santillano et al, "Affect and usage choices in simulation problem solving environments," *Artificial Intelligence in Education: Building Technology Rich Learning Contexts that Work*, R. Luckin, K. R. Koedinger, J. Greer (Eds.), pp. 145-152. Amsterdam, The Netherlands: IOS Press, 2007.

[12] N. Matsuda, E. Yarzebinski, V. Keiser, R. Raizada, G. Stylianides, and W. W. Cohen, and K. R. Koedinger, "Learning by teaching SimStudent – An initial classroom baseline study comparing with Cognitive Tutor," In Proceedings of the International Conference on Artificial Intelligence in Education, G. Biswas & S. Bull (Eds.), pp. 213-221, 2011.

[13] N. Matsuda, W. W. Cohen, K. R. Koedinger, V. Keiser, R. Raizada, E. Yarzebinski, S. P. Watson, and G . Stylianides, "Studying the effect of tutor learning using a teachable agent that asks the student tutor for explanations," Proceedings of the International Conference on Digital Game and Intelligent Toy Enhanced Learning (DIGITEL 2012), M. Sugimoto, V. Aleven, Y. S. Chee, and B. F. Manjon (Eds.) pp. 25-32), 2012.

[14] N. Matsuda, E. Yarzebinski, V. Keiser, R. Raizada, G., Stylianides, and K. R. Koedinger. "Motivational factors for learning by teaching: The effect of a competitive game show in a virtual peer-learning environment," Proceedings of International Conference on Intelligent Tutoring Systems, S. Cerri & W. Clancey (Eds.), pp. 101-111, 2012.

[15] N., Matsuda, E. Yarzebinski, V. Keiser, R. Raizada, W. C. William, G. Stylianides, and K. R. Koedinger, "Cognitive anatomy of tutor learning: Lessons learned with SimStudent," *Journal of Educational Psychology*, in press.

[16] J. Cohen, "A coefficient of agreement for nominal scales, *Educational and Psychological Measurement,* 20, pp. 37-46, 1960.