The Effects of an Interactive Software Agent on Student Affective Dynamics while Using an Intelligent Tutoring System

Ma. Mercedes T. Rodrigo, Ryan S.J.d. Baker, Jenilyn Agapito, Julieta Nabos, Ma. Concepcion Repalam, Salvador S. Reyes, Jr., (*Student Member, IEEE),* Maria Ofelia C.Z. San Pedro

Abstract— We study the affective states exhibited by students using an intelligent tutoring system for Scatterplots with and without an interactive software agent, Scooter the Tutor. Scooter the Tutor had been previously shown to lead to improved learning outcomes as compared to the same tutoring system without Scooter. We found that affective states and transitions between affective states were very similar among students in both conditions. With the exception of the "neutral state", no affective state occurred significantly more in one condition over the other. Boredom, confusion, and engaged concentration persisted in both conditions, representing both "virtuous cycles" and "vicious cycles" that did not appear to differ by condition. These findings imply that – although Scooter is well-liked by students, and improves student learning outcomes relative to the original tutor – Scooter does not have a large effect on students' affective states or their dynamics.

Index Terms— affective dynamics, gaming the system, intelligent tutoring system, Embodied Conversational Agent, Pedagogical Agent

1 Introduction

Increasingly, student interaction with educational software is mediated through interactive software agents of various types. One key form of interactive software agent is the embodied conversational agent (ECA). Embodied conversational agents are the software implementation of the human face-to-face communication metaphor. They are animated anthropomorphisms capable of mixed initiative, verbal and nonverbal communication, and rules for transfer of control [1]. There have also been many recent interactive software agents (also called "pedagogical agents") in the domain of education, which possess all of the attributes of ECAs, except that they do not accept natural language input from students [2], [3], [4], [5], [6], [7], [8], [9]. Instead, human-to-agent communication occurs through menus, or actions in an environment. However, while ECAs and other interactive software agents may be inspired by human-human interaction, and human beings respond socially to computers in many fashions [e.g., 10], humans do not always respond in the same way to a computer as to a human, and it is not fully understood how humans respond to behavior by agents. One key difference between human reactions to agents and human reactions to other humans is that humans tend to react more strongly to assertive and aggressive behavior on the part of other humans than computers, responding less negatively to computers when they behave in this fashion [11]. An extreme example of this is found in evidence that many students respond with delight to being verbally insulted by an ECA embedded in an intelligent tutoring system (personal communication, Sidney D'Mello), a very different response occurs when teachers insult students [12]. As such, it is important to study how humans respond to the behavior of agents, particularly when an agent requests certain behavior on the part of a user or student.

When applied to educational software such as intelligent tutoring systems, agents frequently track student cognition, behavior, or affect in order to provide students with specific support based on individual differences along these dimensions [2], [3], [4], [15]. As such, agent behavior and responses can be considered a type of formative feedback to students [15], and agents often implement a variety of formative feedback strategies. Some of the behaviors which agents manifest in response to student individual differences include the use of emotional expressions [3], [4], non-verbal gestures and communication [5], [16], pedagogical messages [17], requests to stop undesired behavior [3], offering alternate learning experiences [3], and attributional, meta-cognitive, or motivational messages [3], [4]. Interactive software agents have been shown in several studies to positively influence student learning, attitudes, and engagement [2], [3], [4], [5], [6], [7], [16], [17]. However, it is not entirely clear how agents impact students' cognition, behavior, and affect to produce those benefits.

Within this paper, we focus our analyses on the impact of an agent on students' affect, in specific an agent which makes requests of students and responds emotionally to

- *M.M.T. Rodrigo, J. Agapito, J. Nabos, M.C. Repalam, S. S. Reyes, Jr. and M.O.C.Z San Pedro are with Department of Information Systems and Computer Science of the Ateneo de Manila University, Philippines. E-mail: mrodrigo@ateneo.edu.*
- *R.S.J.d. Baker is at Worcester Polytechnic Institute, Worcester, MA. E-mail: rsbaker@wpi.edu.*

Manuscript received (insert date of submission if desired). Please note that all acknowledgments should be placed at the end of the paper, before the bibliography.

their behavior. We hypothesize that a key potential channel through which agents may benefit learners is by changing students' patterns of affect, in particular by increasing engaged concentration and reducing confusion and frustration. We also hypothesize that, although this agent makes assertive requests to the student, this agent's requests will not lead to increased frustration or other negative affect [cf. 11].

Improved affect has been shown to have several positive benefits on learners. In particular, positive affect promotes greater cognitive flexibility and opens the learner to exploration of new ideas and possibilities [18]. Positive affect may also lead to increased situational interest [19], which in turn has been theorized to lead to greater long-term personal interest in the content domain [20]. Negative affective states, on the other hand, can have several negative consequences. For example, boredom has been shown to be correlated with less use of self-regulation and cognitive strategies for learning [21], as well as increases in disengaged and disruptive behavior in class [22]. In addition, frustration can lead to disengagement from a learning task and therefore reduced learning overall [23].

There has been recent attention to affective dynamics, or natural shifts in learners' affect over time [24], [25], [26], building on an early theory in affective computing which postulated that specific affective transitions are likely to be significantly more common than chance [27]. Towards validating, refining, or disconfirming these theories, research in affective dynamics has attempted to determine which affective states tend to persist; which transitions, given a current state, are most likely to occur; and which states tend to lead to a learning or non-learning behavior. The combination of these analyses has led to the discovery of "virtuous cycles" where affective states associated with positive learning (such as engaged concentration) persist, and "vicious cycles" where affective states associated with poorer learning and ineffective learning strategies (such as boredom) persist [8], [24], [25]. Research of this nature can help to model the richness of learners' affective dynamics, and can shed light on key theoretical questions in the field.

In addition, affective dynamics research has the potential to help improve the understanding of how (and whether) interactive software agents influence student affect at the moment-to-moment level. A recent study by Rodrigo et al. [9] examined the differences in affective dynamics promoted by an agent, Paul, within students using Ecolab and M-Ecolab [15], two versions of the same ecology tutor. In terms of cognitive content and pedagogy, the two environments were exactly the same. The principal difference was that M-Ecolab incorporated motivational scaffolding through a virtual learning companion named Paul. Paul's behaviors were driven by a model of the learner's motivation. For example, if a low state of motivation was detected, Paul would use a worried facial expression and the spoken feedback would say: "You're doing well but now try to do even more actions within the activity and if you make an error try again to do the correct action!" Ecolab, by contrast, did not provide the student with any motivational scaffolding. Previous research [15] showed that using the version of the tutor software including Paul led to higher learning gains. Rodrigo et al. [9] found that Paul succeeded in maintaining students' delight over time better than the environment without an agent—a somewhat surprising result, since maintaining delight was not a design goal for Paul. However, Paul did not succeed in maintaining engaged concentration better than the control condition. In addition, Paul did not succeed in either introducing new virtuous cycles or disrupting vicious cycles. Hence, Paul had fairly little impact on students' moment-to-moment affect, suggesting that the learning benefits were produced by some factor other than improved affect.

In this paper, we study the effects of a different agent on the affective dynamics of students. This agent, Scooter the Tutor [3], focuses on reducing the incidence and impact of a specific student strategy associated with poorer learning, gaming the system [28], [29]. Gaming the system consists of the attempt to solve problems and to progress in a curriculum by exploiting the software's help or feedback rather than thinking through the material, for instance through systematic guessing or repeatedly requesting hints at high speed until obtaining the answer. Gaming the system is known to be closely intertwined with boredom [24], [30] – boredom both precedes and follows gaming behavior – suggesting that a system that successfully reduces gaming may reduce boredom as well.

Scooter the Tutor was added to a Cognitive Tutor for Scatterplots, which had previously been shown to lead to large learning gains [31]. Scooter responded to gaming behavior with a combination of meta-cognitive messages (including requests to stop gaming), expressions of positive and negative emotion, and supplementary exercises covering the material the student bypassed through gaming (greater detail on Scooter is given in the next section). The agent successfully reduced gaming and significantly improved gaming students' learning [3] relative to the original tutor.

Scooter provides an interesting opportunity to study the interplay between meta-cognition, affect, and learning. Scooter enforces a positive meta-cognitive pattern that results in positive learning. Understanding whether this positive meta-cognitive pattern and positive learning are associated with changes in moment-to-moment affect may help us understand the role affect plays in behavioral and meta-cognitive changes produced by agents.

Hence, within this paper, we study student affect while using Scooter in a fine-grained fashion, focusing on the dynamics of affect when using Scooter and when using the same tutor without Scooter. Does Scooter create positive or negative affect where it was not previously present? Will Scooter disrupt students' boredom and frustration, perhaps ending vicious cycles? Contrastingly, will Scooter disrupt students' engaged concentration, perhaps ending virtuous cycles? Does Scooter turn boredom and confusion (the affective states that most precede gaming behaviors [24]) into engaged concentration? Or more negatively, into frustration? What other impacts on student affect does Scooter have?

2 TUTOR AND AGENT

Scooter the Tutor, the interactive software agent we study in this paper, was implemented in the context of the Scatterplot Tutor [3], an intelligent tutoring system that teaches students how to create and interpret scatterplots of data. The Scatterplot Tutor was originally designed as part of the Middle School Mathematics Tutor [32]. Problems in the Scatterplot tutor involve a range of domains, from parcel delivery, to patents and medical discovery, to financial decision-making.

An example of a scenario that a student must solve is:

> *Samantha is trying to find out what brand of dog food her dog food her dog Champ likes best. Each day, she feeds him a different brand and sees how many bowls he eats. But then her mom says that maybe her dog just eats more on days when he exercises more.*
>
> *Please draw a scatterplot to show how many bowls the dog eats, given the dog's level of exercise that day.*

Figure 1 has the worksheet with the data that the student must plot.

Figure 2 is the variable type tool. Using the tool, the student must identify the nature of the data available to him or her and determine whether it is appropriate or inappropriate for a scatterplot.

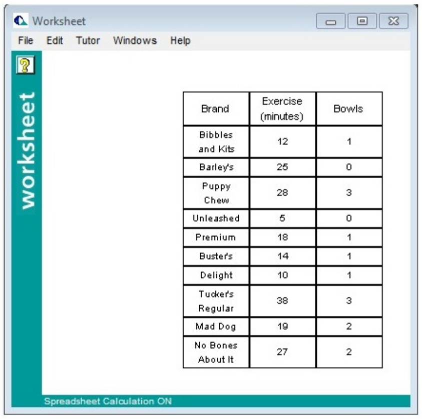
Once the student completes the variable type tool, the graphing area appears (Figure 3). This is where the student must construct the scatteplot.

To help the students determine the scales of the x and y axis, they use the scaling tool (Figure 4).

By answering the questions on the scaling tool, students determine the correct starting value of each axis and the appropriate increments. Each action a student takes when using the software is associated with one or more component skills that, when attained, lead to the mastery of the topic. To help students solve the problems, the tutor provides step-by-step guidance such as contextual hints about what to do next, feedback on correctness, and just-in-time messages for common errors.

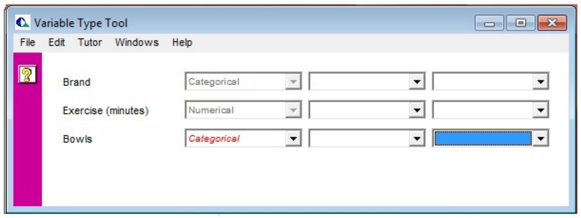
Baker et al. developed an experimental version of the Scatterplot Tutor with software agent named "Scooter th Tutor" [3] (Figures 5 and 6), using graphics from the Microsoft Office Assistant.

Scooter was designed to both reduce the incentive to game the system, and to help students learn the material that they were avoiding by gaming, while affecting non-gaming students as minimally as possible.



Brand	Exercise (minutes)	Bowls
Bibbles and Kits	12	1
Barley's	25	0
Puppy Chew	28	3
Unleashed	5	0
Premium	18	1
Buster's	14	1
Delight	10	1
Tucker's Regular	38	3
Mad Dog	19	2
No Bones About It	27	2

Fig. 1. Worksheet



Brand: Categorical

Exercise (minutes): Numerical

Bowls: Categorical

Fig. 2.Variable Type Tool

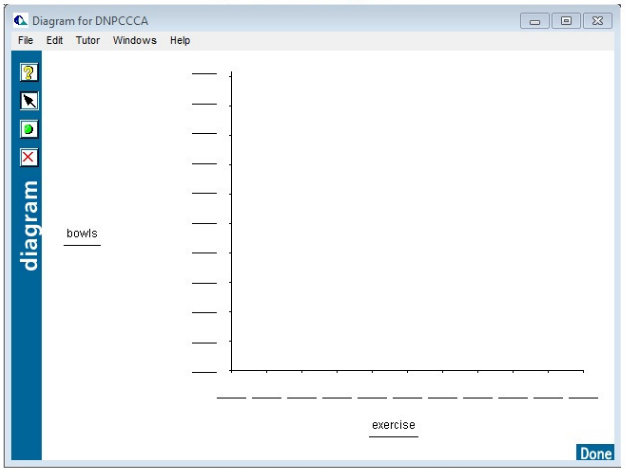
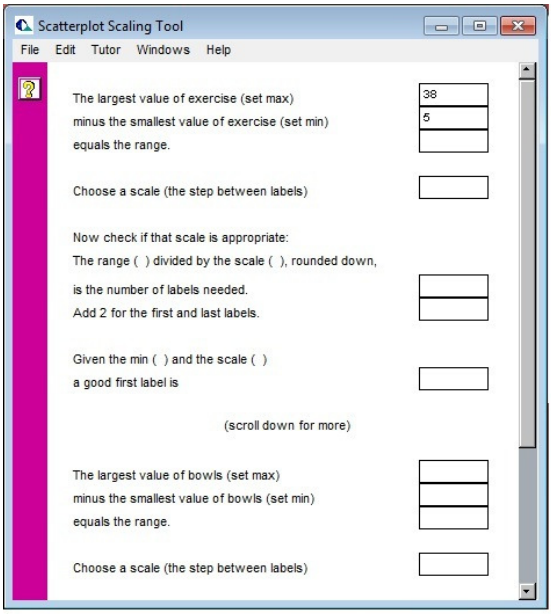


Fig. 3. Scatterplot Work Area



The largest value of exercise (set max) 38

minus the smallest value of exercise (set min) 5

equals the range.

Choose a scale (the step between labels)

Now check if that scale is appropriate:

The range ( ) divided by the scale ( ), rounded down,

is the number of labels needed.

Add 2 for the first and last labels.

Given the min ( ) and the scale ( )

a good first label is

(scroll down for more)

The largest value of bowls (set max)

minus the smallest value of bowls (set min)

equals the range.

Choose a scale (the step between labels)

Fig. 4. Scaling Tool

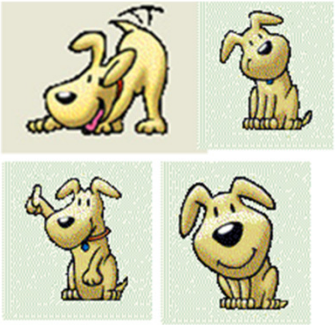


Fig. 5. Happy expressions of Scooter

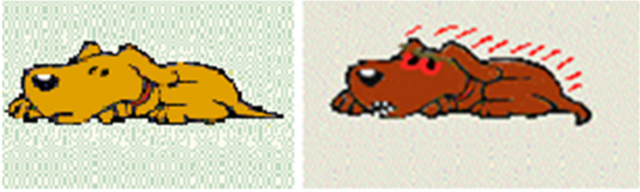
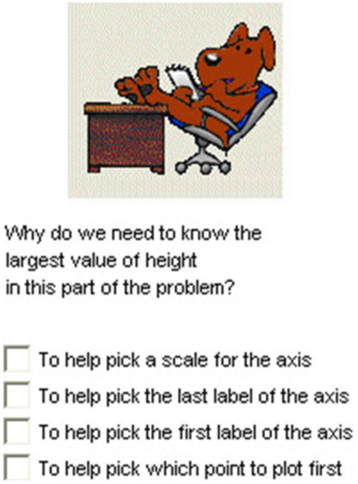


Fig. 6. Sad (left) and Angry (right) expressions of Scooter

When a student is not gaming, Scooter looks happy and occasionally gives the student positive messages (Figure 5). Scooter's behavior changes when the student is detected to be gaming the system, using a machine-learned detector of gaming the system [33]. If the detector assesses that the student has been gaming, and the student has not yet obtained the answer, Scooter displays increasing levels of displeasure (starting with sadness, and terminating in expressions of anger, shown in Figure 6), to signal to the student that he or she should now stop gaming, and try to get the answer in a more appropriate fashion. These expressions of emotion are combined with simple meta-cognitive messages, suggesting that the student should work carefully in order to learn. If the student obtains a correct answer through gaming, Scooter gives the student a set of alternate supplementary exercises designed to give the student another chance to cover the material that the student bypassed by gaming this step, concluding the set with a reminder that the student should not game (Figure 7).

The detector of gaming [33] was developed to predict assessments of gaming the system produced through quantitative field observations (detail on this method is given later in this paper). This detector has been validated to be over 80% correct at distinguishing gaming and non-gaming students, even when applied to entirely new students or entirely new tutor lessons [33]. Detector inaccuracy is roughly evenly distributed between false positives and false negatives, implying that some students will receive unnecessary supplementary exercises, and some students will not receive exercises that would have been appropriate. However, students who receive unnecessary exercises are not significantly disadvantaged; students who know the relevant skill can answer the question quickly and resume working, and students who do not know the skill are likely to benefit from the exercises in the same fashion as gaming students. As such, the interventions can be considered "fail-soft," having low consequences for false positives. The goal in designing Scooter was to benefit students in three fashions. First, by representing how much each student had been gaming, Scooter both serves as a continual reminder that the student should not game, and lets teachers know which students were gaming recently.



Why do we need to know the
largest value of height
in this part of the problem?

□ To help pick a scale for the axis
□ To help pick the last label of the axis
□ To help pick the first label of the axis
□ To help pick which point to plot first

Figure 7. Scooter giving supplementary exercises.

Second, Scooter was intended to invoke social norms in students (cf. [10]) by expressing negative emotion when students game. Scooter's display of anger is a natural social behavior in this context; if a student systematically guessed every number from 1 to 38 when working with a human tutor, it seems reasonable to expect that the human tutor would become impatient or upset (and in our experience from classroom observation, teachers do indeed become impatient and upset when they see students gaming). As discussed earlier, human responses to software agents are not identical to their responses to identical behavior coming from humans [11]. Thus, it is likely that students will not respond in the exact same fashion to an angry cartoon dog as to an angry human being; in specific, the findings of [11] suggest that students should be less upset by the angry cartoon dog than they would by a human, potentially making the negative effects of this form of feedback less likely to lead to negative affect.

Third, by giving students supplemental exercises targeted to the material the student was gaming through, Scooter gives students a second chance and another way to learn material he or she may otherwise miss entirely. Additionally, supplemental exercises may change the incentive to game – whereas gaming might previously have been seen as a way to avoid work, it now leads to extra work.

When tested in classrooms in the USA, Scooter successfully reduced gaming and significantly improved gaming students' learning [3]. However, students who received many interventions from Scooter (e.g., students who gamed the system) liked him significantly less than students who received few or no interventions [31].

Hence, the very students who were benefitted by Scooter were the ones who disliked him. Within this paper, we study Scooter's impact on student affect and affect dynamics.

There was reason to hypothesize that Scooter would have a positive impact on students in the Philippines, perhaps an even greater impact than on students from the USA. Adherence to social norms is one of the characteristics of Philippine culture. By contrast, inter-cultural research has suggested that American culture values independence over obedience [34], [35]. Respect for authority is less important than in other cultures [34], [35] and happiness is seen as resulting from personal achievement [36]. Philippine society is characterized as one in which obedience and respect for authority are considered more important than independence and self-direction [34], [35]. [37] observed that behaving in a socially acceptable fashion and avoiding all outward signs of conflict is more important in the Philippines than in the USA. One caveat of this cultural characteristic is that people from the Philippines and other Asian societies tend to mask their individual indiosyncracies behind the persona associated with one's role in society [38]. They may engage in behaviors that, on the outside, appear to conform to the social norm, but still flout authority in less overt ways, a phenomenon that [38] labels as "half ungestured feelings, half unfelt gestures." Considering that gaming is fairly surreptitious behavior, gaming may therefore be higher in the Philippines than in the USA.

3 METHODS

We conducted the study in a large, urban high school in Quezon City (pop. 2.7 million) in the Philippines. As of 2008, the school had 5,368 students, predominantly Filipino, and 216 teachers [39]. The school's community is relatively poor. A survey of these childrens' parents revealed that about one-half of respondents were unemployed and approximately 70% of households earned PhP10,000 per month or less (approximately US$230.00) [29]. The school had 32-bit Windows XP computers intended for student use, however many were in disrepair or were kept in storage and not used for instruction.

The study design was reviewed and approved by the Ateneo de Manila University's Ethics Committee. We met the principal of the participating school to explain the goals, materials, and methods of the study. Both the principal and Mathematics Subject Area Coordinator of the participating school gave permission to us to conduct the study. Finally, we wrote the parents of selected high school freshmen, inviting their children to participate in the study. In the letter, we explained what activities they were going to perform. We also explained that we would observe them and record their work. They gave us written permission to involve their children in study.

Participating students' ages ranged from approximately 12 to 14. We collected complete sets of data from a total of 126 students, with 64 in the experimental condition, and 62 in the control condition. A faulty USB drive corrupted log data from some of the students causing data loss for four students.

Students were assembled into groups of 10, due to the size of the school's computer lab. In three cases, absenteeism led to groups of 9 members each. Each student in each group answered a content-related pre-test together with a motivationnal questionnaire. Afterwards, students viewed conceptual instruction on scatterplot generation and interpretation, delivered via a PowerPoint presentation with voiceover and some simple animations. The students then used the Scatterplot Tutor for 80 minutes. Finally, the students took a post-test. The students in the experimental condition also answered a questionnaire regarding their attitudes towards Scooter.

The tests and questionnaires were exactly the same questionnaires used in previous research on Scooter in the USA [31], with one exception – pre-test items asking students for their thoughts on intelligent tutoring systems were omitted, since these students did not have prior experience with intelligent tutoring systems. Many of the items were adapted from past questionnaires on student responses to educational software, teachers, and intelligent agents (cf. [40], [41]). The questionnaires were expressed as Likert scales, with responses possible between 1 (Strongly Disagree) and 6 (Strongly Agree). Items used are shown in Table 1. (One item accidentally made reference to other software not used by this population, and will be omitted from analysis).

TABLE 1
SCOOTER (POST-TEST) QUESTIONNAIRE ITEMS

1. Scooter is friendly
2. Scooter is smart
3. Scooter treats people as individuals
4. Scooter ignores my feelings
5. I feel that Scooter, in his own unique way, genuinely cares about my learning
6. Scooter wants me to do well in class
7. Scooter is irritable

While students were using the software, we collected data on each student's pattern of affect during tutor usage, using a quantitative field observation method originally proposed in [21], and refined and used within [9], [24], [30] and [42].

Quantitative field observations are just one of several methods used to collect affect data on subjects, and there has been considerable debate as to which method (if any) should be considered the "gold standard" [43], [44]. Other methods include the use of video annotation [43], screen replay annotation [45], automated detection using sensors (cf. [46], [47]), "emote-aloud"and retrospective "emote-aloud" methods [48], and self-report through a user interface during the learning task [13] [49]. There are several factors to consider when selecting which method to use to code affect. In particular, many standard methods are difficult to use in authentic learning settings such as high school classrooms. Video annotation takes considerable time, and gives a more limited range of information on affect than physical presence, as context is difficult to infer and posture assessment is more difficult. For this reason, video annotation of affect typically focuses on facial expressions. However, facial coding of complex affect conducted in this fashion often achieves poor Kappa, even with highly-trained coders (e.g., [50]). In addition, within classrooms, students frequently leave their seats and/or may not be looking at their screens when seeking help from their teacher or giving help to another student, frequent behaviors in classroom use of intelligent tutoring systems (e.g., [28], [51]). During these periods of time, video annotation of students' faces may be infeasible. The use of sensors such as biometrics, eye-tracking, and posture tracking require specialized equipment and is difficult to scale. Though researchers have recently begun to use sensors for classroom research on affect (e.g., [49]), validating sensor-based models for this context remains a significant research challenge, and breakage and loss of equipment is a significant problem (personal communication, Ivon Arroyo). In addition, many sensors, such as posture chairs, are not useful if the student is out of his or her seat (or even sitting differently in order to talk to another person). Emote-aloud methods are similarly difficult to use in a classroom at scale. When twenty or thirty students are emoting aloud at the same time, audio interference is a significant problem. In addition, emoting-aloud is likely to interfere with normal collaborative and help-seeking behaviors which students use in classroom settings (e.g., [28], [51]). Self-report given through a user interface is more feasible, and has been frequently used in classrooms (e.g., [13], [49]), but carries challenges of its own; in particular, self-report of affect can disrupt student concentration and even change affect (e.g., it can annoy students), if requested in the middle of a problem. Self-report requested between problems may be less useful for studying fine-grained affect during learning, as current self-report methods are not able to capture the sequence of affect while solving a problem. Another method, self-report at a time of the student's choosing, may not capture all changes in affect; a confused or intensely concentrating student may not take the time to indicate their affect in an interface. In addition, self-report measures given during problems may be ignored when a student is talking to another person, delaying response and potentially missing key affect.

For these reasons, we adopt the method of quantitative field observations [9], [24], [28], [30], [42], [52], [53] for observing student affect. Field observers have access to multiple channels of information; they can look at the student's work context, actions, utterances, facial expressions, and body language. These are the sorts of cues used by human beings when assessing affect in real life; typically humans use multiple cues in concert for maximum accuracy, rather than attempting to select individual cues [54]. Past observational research studying affect in context has used this set of cues to code affect with high reliability [55], also achieving good agreement with self-reports. As such, field observations have the potential to be more accurate than video observations, where less information is available to coders. In addition, field observations are robust to changes in student location and student conversations, unlike other methods.

We do not suggest that this technique is without limitations. Perhaps the biggest challenge to the use of field observations is that there is no way to revisit and re-code the data. Video data can be re-analyzed using alternate coding schemes or by a second research group, providing greater potential for validating and using the data. In addition, the observation process is a tiring experience which requires the observer's full attention. The accuracy of the data can be influenced by observer fatigue, as well as by distraction. Training in the quantitative field observation process is time-consuming, and a minority of individuals are not capable of viewing study participants at the optimal angles for this method. Finally, we have informally found that observations are substantially less accurate if the observer is from a different national background than the participants being observed. Hence, attention to these sorts of issues must occur in order for the method to be used effectively. Nonetheless, for this specific study, quantitative field observations was judged to be the best method, for the reasons discussed earlier.

Our pool of observers was composed of three of this paper's co-authors, all of whom have prior classroom experience as teachers. The observers trained for the task through a series of pre-observation discussions on the meaning of the affect and behavior categories, oriented around a coding manual developed in prior research (more detail on this coding manual is given in [24], [30]). Observations were conducted according to a guide that gave examples of actions, utterances, facial expressions, or body language that would imply an affective state. After these discussions, additional field-based training was conducted prior to the actual study; the observers trained with the first author at the same school, repeatedly conducting and discussing observations, and then conducting independent rounds of observation. These observations were then summarized into a confusion matrix, which was studied and discussed further to resolve remaining discrepancies. Afterwards, a formal check of inter-rater reliability was conducted.

During the data gathering sessions, two of the three trained observers coded each student's affective state in real-time. The observers conducted 24 observations per student. Each student therefore had a total of 48 observations. In order to avoid bias towards more interesting or dramatic events, the coders observed the students in a specific order determined before the class began. Any affective state by a student other than the student currently being observed was not coded. Each observation lasted for 20 seconds. Each student was observed once every 200 seconds (e.g., 180 seconds between observations), an interval determined by the number of students being observed and the number of observers. Coders were synchronized using a timed PowerPoint presentation. If two distinct affective states were seen during an observation, only the first state observed was coded. In order to avoid affecting the current student's affect if they became aware they were being observed, the observers viewed the student out of peripheral vision, with quick glances, or at an angle while appearing to look at another student. Conducting coding in this manner requires training and practice, but in our experience (having now trained approximately 30 coders between the first and second authors), the majority of people learning this method have been able to learn to code in this fashion within a relatively short time (and those who cannot do so have been able to clearly identify that fact as soon as they being coding).

Following the work of [24], [25], [26], [30], [42], our coding scheme consisted of seven categories: boredom, confusion, delight, engaged concentration, frustration, neutral and surprise. Some of these states have also been referred to as cognitive-affective states (e.g., [24], [25]) as they stem from cognitive events (in the case of confusion, not understanding something) even if they are experienced as affect. It has been argued that these affective states are of particular importance for research on affect during learning, and are more representative of student affect during learning than the basic emotions of anger, fear, sadness, happiness, surprise, and disgust [48], [56]. It is worth noting that in previous research, engaged concentration has been referred to as "flow" (cf. [26], [9]). We call it "engaged concentration" instead, as it is not clear whether the momentary affective states typically associated with the complex construct of flow (cf. [57]) are accurate indicators of that more comprehensive construct.

The observers' inter-rater reliability was then tested using Cohen's Kappa. Kappa was found to be 0.68. This level of Kappa is considered to be substantial though not excellent and is comparable to or better than Kappa values previously reported in other published papers where judgments are made of genuine and naturally occurring expressions of affect and emotion. For example, past research where experts made judgments of learners' affect from video taken in laboratory settings has involved Kappa values between 0.3-0.4 (e.g. [43], [61]). Prior field observational research on affect has involved Kappa values between 0.6 and 0.75 (e.g. [9], [24]). In general, the level of Kappa reported here indicates that measurement of affect will have some noise. Table 2 shows that the primary disagreements between coders were in terms of distinguishing confusion from engaged concentration and, to a lesser extent, boredom. It is not yet fully understood which affective states are routinely confused by humans; this finding suggests it may be worth investigating difficulties in coding confusion further. This limitation should be taken into account when considering whether the limitations of field observations or self-report are preferable.

4 Results

4.1 Learning

In both conditions, as in [3], students had statistically significant improvement from pre-test (M=17%, SD= 27%) to post-test (M=61%, SD= 31%), t(121)=13.95, two-tailed p<0.001, Cohen's *d*=1.53, for a paired t-test. Also as in [3], there was no main effect for learning between conditions, t(120)= 0.90, two-tailed p=0.19, Cohen's *d*= -0.47, for a two-sample t-test assuming equal variances, though the trend appeared to be in favor of the control condition. Just as in the earlier USA study of Scooter, only a minority of students game the system, and gaming detection is not perfect (it can be expected to be 80% accurate at distinguishing gaming students from non-gaming students [33]). As such, differences in learning in the overall population may not be indicative of effects on the sub-population receiving interventions.

TABLE 2
Confusion Matrix for Affect Observatons

	BOR	CON	DEL	ENG	FRU	N	SUR
BOR	152	32	1	10		4	
CON	23	1246	3	164	2	8	1
DEL	1	8	29	4		1	1
ENG	5	263	2	986	1	5	1
FRU		3	1		7		
N	1	2		1		29	
SUR							2

BOR = boredom; CON = confusion; DEL = delight; ENG = engaged concentration; FRU = frustration; SUR = surprise; NEU = neutrality.

In past work [e.g. 3], Scooter's impact on learning was studied in greater depth by analyzing the relationship between the number of supplementary exercises (e.g., primary gaming interventions) received and learning in the experimental condition, and the relationship between gaming and learning in the control condition. In that previous work, students who received the most exercises from Scooter (top third, in terms of number of exercises received) had higher learning gains than students who received fewer exercises, a major shift from the control condition (and previous research on gaming [e.g., 28]), where gaming was associated with poorer learning. In the data collected for this paper, there was considerable variance in the number of exercises received by students, with an average of 3.98 exercises and a standard deviation of 3.01 exercises. There was the appearance of a trend towards differences in learning, by the number of exercises received. Students who received moderate numbers of exercises from Scooter (middle third in terms of number of exercises received – 2 to 4 exercises) appeared to demonstrate better learning, an average pre-post gain of 52.2 points, than those who received the most exercises (top third of exercises received – 5 or more exercises), with an average gain of 36.4 points, or those who received the fewest exercises (bottom third of exercises received – 0 or 1 exercises), with an average gain of 31.1 points. However, the overall difference between groups was not statistically significant, F(2, 56)=2.13, p=0.13, $\eta^2 = 0.07$, for a one-way ANOVA. Hence, differences between individual groups will not be presented here. Overall, the correlation between interventions received and learning was a very small -0.06, which is not significantly different than chance, t(58)=0.45, p=0.65, for a test of the statistical significance of a correlation coefficient.

However, despite the failure to replicate the experimental condition result previously seen in [3], in the control condition, gaming (assessed by the same detector used to drive interventions in the experimental condition) remained associated with poorer learning. In a regression predicting post-test score, a student's percentage of time spent gaming the system was negatively associated with the post-test, r=-0.45, t(61)= -3.34, two-tailed p=0.001, for a test of the significance of a correlation coefficient. Gaming and pre-post gain were marginally negatively correlated, r= -0.225, t(61)= -1.81, two-tailed p=0.08, for a test of the significance of a correlation coefficient.

As such, it appears that the negative relationship between gaming and learning may persist in the control condition, but there is not the appearance of a similar relationship between the interventions (driven by gaming detected in the same fashion) and pre-post gains in the experimental condition. This result does not replicate the earlier result in [3], where substantial numbers of interventions were actually associated with better learning. The difference in correlation between gaming and pre-post gains was not significantly different between conditions, Z=0.92, p=0.36, for a test of the significance of the difference between two correlation coefficients for independent samples, using the Fisher Z transformation. As such, the question of whether there is a difference in learning due to Scooter's interventions or the presence of Scooter, within this study, can not be conclusively answered.

4.2 OVERALL ATTITUDES AND AFFECT

In past research in the United States, students have generally reported disliking Scooter [31], particularly among students who had improved learning. Students in the current study, however, reported liking Scooter. In the post-test questionnaire regarding their attitudes towards Scooter, the students said that he was friendly (M= 4.65, 95% CI 4.26-5.04), smart (M=5.14, 95% CI 4.86-5.42), that he treated them as individuals (M=4.67, 95% CI 4.34-4.99), that he did not ignore their feelings (M=2.48, 95% CI 2.08-2.87), that he genuinely cared about them (M=5.55, 95% CI=5.34-5.76), that he wanted them to succeed in class (M=5.23, 95% CI = 4.90-5.56) and that he was not irritable (M=3.01, 95% CI = 2.57-3.46). None of these 95% confidence interval ranges cross the interval mid-point (3.5); if the 95% confidence interval crossed the interval mid-point for a construct, it would imply that it was uncertain whether student responses were overall positive or negative for that constructs. Since none do, it can be concluded that students' attitudes towards Scooter were on the whole positive. In addition, there is no evidence that students who received interventions from Scooter liked Scooter more or less than other students. To this end, we computed correlations between the number of interventions (supplementary exercises) each student received from Scooter, and responses on each of the questionnaire items about Scooter. None of these correlations were statistically significant.

Somewhat surprisingly, and unlike prior results for Scooter [e.g., 3], gaming actually was higher for the experimental condition (M=18.8%, SD=4.4%) than the control condition (M=12.0%, SD=2.7%), a statistically significant difference, t(120)= -1.99, two-tailed p=0.05, Cohen's *d*=1.86, for a two-sample t-test assuming equal variances. One qualitative finding that may explain this behavior is that many students reported intentionally gaming in order to receive Scooter's interventions. Students reported treating Scooter's interventions as help, a means of understanding the subject matter better. They did not regard the supplementary exercises as penalties.

Another test of the impact of Scooter, and Scooter's interventions, is the overall proportion of each affective state, in each condition. If students disliked Scooter and his interventions, there should be evidence for more negative affect--such as frustration or boredom—in the experimental condition. Correspondingly, if students liked Scooter and his interventions, there should be evidence for more positive affect—such as delight or engaged concentration—in the experimental condition.

The occurrence of each affective state is shown in Table 3. The first number represents the number of times that the given affective state was observed. The second number represents the percentage of observations where each affective state was recorded. In these computations, we included all observations, even those in which the two observers disagreed. In cases of disagreement, each coding was given half-credit. Eliminating cases where disagreement occurs might systematically bias the sample against affective states that are more likely to be the subject of disagreement, a particularly important consideration given the presence of coder disagreement in terms of confusion. Note that coder error sometimes resulted in no observation for a particular 20-second period.

TABLE 3
PERCENTAGE OF OCCURENCE FOR EACH AFFECTIVE STATE

Affective State	Control	Experimental
Boredom	179 6.01%	204 6.64%
Confusion	1,406 47.24%	1,605 52.25%
Delight	39 1.31%	42 1.37%
Engaged Concentration	1,293 43.45%	1,142 37.17%
Frustration	13 0.44%	8 0.26%
Surprise	4 0.13%	3 0.09%
Neutral	26 0.87%	55 1.79%
None (Coder error)	16 0. 54%	13 0.42%

Confusion was the most frequent affective state in both conditions. It occurred 47.24% of the time in the control group, and 52.25% of the time in the experimental group, an apparent difference which was not statistically significant, t(124)=-1.38, two-tailed p = 0.17, Cohen's *d*=0.25, for a two-sample t-test assuming equal variances. The second most common affective state was engaged concentration, occurring 43.45% of the time in the control group and 37.17% in the experimental group; the apparent difference between conditions was not statistically significant, t(124)=1.52, two-tailed p=0.13, Cohen's *d*=0.27, for a two-sample t-test assuming equal variances. The third most common affective state was boredom, occurring 6.01% of the time in the control group and 6.64% of the time in the experimental group, a difference which was not statistically significant, t(124)=-0.24, two-tailed p=0.81, Cohen's *d*=0.04, for a two-sample t-test assuming equal variances. All the other affective states occurred under 2% of the time in both conditions.

Interestingly, confusion seemed to be a more prevalent part of students' experiences with the Scatterplot Tutor (in both conditions) than other intelligent tutoring systems. Compared to other studies on affect using this method and similar populations, students using the Scatterplot tutor exhibited substantially more confusion than students using Ecolab(12.7%) [9], M-Ecolab (12.9%) [9], The Incredible Machine (11%) [30], or Aplusix (13%) [42]. Correspondingly, engaged concentration was less common in both conditions than in prior systems, including Aplusix (68%), The Incredible Machine (62%), Ecolab (61.5%) and M-Ecolab (67.4%). Boredom was seen less frequently than in many previous studies with the same method and similar population, including past studies of Ecolab (15.2%), M-Ecolab (12%), and The Incredible Machine (7%), though higher than Aplusix (3%).

Hence, overall, it does not appear that the broad patterns of affect differed in significant ways between the two versions of the Scatterplot Tutor. But it appears that the Scatterplot Tutor generally was a difficult environment for students, one where they were frequently confused but rarerly bored. Fortunately, despite the high degree of confusion, frustration was quite rare.

4.3 PERSISTENCE OF AFFECTIVE STATES

Beyond the overall prevalence of each state, it is important to consider how Scooter influenced the persistence of each state within the Scatterplot Tutor. For example, it may be that Scooter disrupted "vicious cycles" of student frustration or confusion [26], even if the overall prevalence did not significantly reduce. Hence, we look within each condition at how persistent each state was – e.g., the probability of a student being in the same specific affective state or a different state 180 seconds later (the time between two observations of the same student within our protocol). In computing the likelihood of an affective transition, it is important to take into account the base rates of each affective category. Confusion was the most frequent affective state within both systems; thus, confusion is likely to be the most common affective state that follows any other affective state. Hence, we use D'Mello's [25] transition likelihood metric **L** in order to appropriately account for the base rate of each affective category in assessing how likely a transition is. D'Mello et al's **L** [25] gives the probability that a transition between two affective states will occur, given the base frequency of the destination state, and is computed:

$$L = \frac{\Pr(NEXT \mid PREV) - \Pr(NEXT)}{(1 - \Pr(Next))} \qquad (1)$$

L is scaled between -∞ to 1. A value of 1 means that the transition will always occur. A value of 0 means that the transition's likelihood is exactly what it would be, given only the base frequency of the destination state (i.e., this transition occurs with exactly the frequency that would occur if transitions were random). Values above 0 signify that the transition is more likely than it could be expected to be given only the base frequency of the destination state, and values under 0 signify that the transition is less likely than it could be expected to be, given only the base frequency of the destination state.

In studying the persistence of affective states, we study the transition from a state to itself (e.g., BOR to BOR). To this end, for each self-transition, **L** was calculated for each student and then the mean and standard error across students were obtained. Note that if a student never engages in a specific affective state or always engages in the same affective state, the formula for **L** is undefined, causing the number of students for which **L** can be calculated (and therefore the number of degrees of freedom) to vary by affective state. Given these results, it is possible to determine if a given transition is significantly more likely than chance, given the base frequency of the next state, using the two-tailed t-test for one sample (all tests in this section are of this form). Table 4 shows the persistence of each affective state for the control and experimental conditions respectively. The first number in each cell represents the mean **L** value while the second number, in parentheses, represents the standard deviation. Cells in grey are statistically significant at p < 0.05.

TABLE 4
TRANSITION BETWEEN AFFECTIVE STATES

	BOR-BOR	CON-CON	DEL-DEL	ENG - ENG	FRU – FRU	SUR – SUR	NEU - NEU
CONT	0.19 (0.35)	0.16 (0.41)	0.04 (0.13)	0.13 (0.43)	0.08 (0.19)	0.00 (0.00)	0.02 (0.12)
EXP	0.25 (0.35)	0.18 (0.39)	0.04 (0.13)	0.09 (0.36)	0.06 (0.13)		0.06 (0.21)

BOR = boredom; CON = confusion; DEL = delight; ENG = engaged concentration; FRU = frustration; SUR = surprise; NEU = neutrality. Transitions that were statistically significantly more common or rare than chance are shaded in grey. Blank cells indicate transitions that were too rare for D'Mello's L to be calculable. Values indicate mean values of D'Mello's L for each transition (e.g., Standard deviations are shown in parentheses.

We find many similarities in the persistence of affect between conditions. In both cases, boredom, confusion and engaged concentration tend to persist, results consistent with prior results [9], [25], [26]. A student who is bored in either condition will tend to stay bored (control: **L**=0.19, t(22)=2.62, two-tailed p=0.02; experimental: **L**=0.25, t(22)=3.49, two-tailed p < 0.01). As mentioned earlier, all statistical tests in this section are two-tailed t-tests for one sample. A student who is confused will tend to stay confused (control: **L**=0.16, t(60)=3.07, two-tailed p<0.01; experimental: **L**=0.18, t(62)=3.65, two-tailed p < 0.01). A student who is engaged will tend to stay engaged (control: **L**=0.13, t(59)=2.37, two-tailed p=0.02; experimental: **L**=0.09, t(61)=2.05, two-tailed p=0.04).

Similarities also existed even among non-persistent states. Although delight was previously found to be persistent in the M-Ecolab environment [9], another environment incorporating an agent, it was not persistent in either of our conditions (control: **L**=0.04, t(12)=1.14, two-tailed p=0.27; experimental: **L**=0.04, t(12)=1.05, two-tailed p = 0.31). Similarly, frustration was not seen to be persistent (control: **L**=0.07, t(4)=1.04, two-tailed p=0.36; experimental: **L**=0.06, t()=1.07, two-tailed p = 0.36).

We compared the mean L values of the self-transitions within the control condition against those of the experimental condition to determine whether differences existed in the degree to which these different states tended to persist (or not). We found no significant differences in the persistence of states between conditions for any of the affective states (Table 5).

TABLE 5
COMPARISON OF MEAN L VALUES FOR CONTROL AND EXPERIMENTAL CONDITIONS

Transitions	Comparison
BOR-BOR	t(46)=-0. 59, two-tailed p=0.56
CON-CON	t(124)=-0.26, two-tailed p=0.80
DEL-DEL	t(26)=0.09, two-tailed p=0.93
ENG-ENG	t(122)=0.56, two-tailed p=0.58
FRU-FRU	t(9)=0.14, two-tailed p=0.89
SUR-SUR	
N-N	t(36)=-0.60, two-tailed p=0.55

BOR = boredom; CON = confusion; DEL = delight; ENG = engaged concentration; FRU = frustration; SUR = surprise; NEU = neutrality. The comparison was made using a t-test assuming equal variance. The blank cell indicates that the transition was too rare for the t-test to be calculable.

The similarity of the patterns of persistence of these states in the two conditions implies that, although students report liking Scooter, Scooter does not appear to disrupt or reinforce either vicious cycles involving boredom or virtuous cycles involving engaged concentration.

5 DISCUSSION, CONCLUSIONS, AND FUTURE WORK

Interactive software agents have been shown in several studies to improve student motivation, engagement, and learning. In this paper, we studied an agent named Scooter the Tutor, embedded into a Cognitive Tutor for Scatterplots. This agent has been shown to statistically significantly improve gaming students' learning in prior research [3], a finding which was not replicated here (though not conclusively). We found that students typically reported liking the agent, but – surprisingly – there were no significant differences in observed affect between the two conditions.

Given some evidence that the agent increased student gaming despite the positive responses about Scooter on the questionnaire, there is evidence that students may have perceived and interacted differently with Scooter in the Philippines than in the USA. The gaming behavior of the students in the Philippines suggest Scooter's interface design did not successsfully leverage Philippine society's preference for outwardly smooth interpersonal relationships. We had anticipated that the cultural preference for outwardly smooth interpersonal relationships, combined with Scooter's irritability when students gamed, might lead students to game less. The opposite occurred. One possible explanation is that students interpreted Scooter differently than expected; given that students in the Philippines did not perceive the (negative-emotion displaying) Scooter as irritable, it becomes open to question whether they perceived that Scooter was displeased by gaming. Instead, students appeared to perceive Scooter solely as a useful, helpful, caring authority figure with expertise in the subject matter. Hence, gaming the system became a way to get more help from Scooter (making this a non-harmful form of gaming [cf. 33]). This perception of Scooter might also account for the fact that students in the Philippines reported liking Scooter, whereas students in the US did not like Scooter.

In future work, it may be useful to expand our post-test questionnaire to explore students' comprehension of Scooter's role in the learning process. It also may be valuable to expand the post-test questionnaire to explore the degree to which students believe Scooter responds in a human-like fashion and provides human tutor-quality responses.

One of the ways in which interactive software agents are hypothesized to influence students is through changing the degree to which students experience certain affective states. However, the data presented here shows no significant difference between the incidence of affective states between conditions. This result replicates findings in prior research involving a motivational agent in an ecology tutor [6], where differences in proportion of affect were not found. In general, these results suggest that influencing moment-to-moment student affect during genuine learning is difficult, even when more general attitudes are influenced.

Interactive software agents are further hypothesized to influence the pattern of student affect during learning. In prior research involving a motivational agent in an ecology tutor [6], there was no difference in the persistence of engaged concentration, boredom, or frustration; there was, however, greater persistence for delight.

The study presented here has relatively similar findings. Both conditions (with and without the agent) had a vicious cycle involving boredom, and a virtuous cycle involving engaged concentration, but there was no evidence that Scooter disrupted or reinforced either vicious cycles involving boredom or virtuous cycles involving engaged concentration. The increased delight seen with the agent in M-Ecolab was not replicated with Scooter.

As such, the evidence of this study suggests that our prior hypothesis as to the effects of this agent may not be correct. While Scooter was well-liked by students in this study, he did not substantially alter student affect or affective dynamics at the moment-to-moment level. This is somewhat surprising, since the key behavior that Scooter was designed to address – gaming the system – was previously found to be closely intertwined with boredom [24].

The pattern of results presented is suprising for a second reason – the students who reported liking Scooter a great deal nonetheless experienced no less boredom or frustration, and no more delight and engaged concentration, than students who did not like Scooter. Hence, the students' reasons for liking Scooter must not be due to moment-to-moment improvements in student affect. This raises the question of what factors explain students' liking of Scooter. Software agents are rare in the Philippines—it is possible that the students' positive attitudes were essentially just a novelty effect. Another possibility is that students liked Scooter because he was educationally helpful, not because he improved their affect. In other words, students liked Scooter because he was helpful and therefore rated him more highly in other areas as well. A corresponding effect is seen in [16], who found that students rated an agent with pedagogical feedback higher on dimensions that were unrelated to the feedback given, such as the agent's ability to recognize and interpret the student's utterances. However, one factor seen in both this study and the previous study of Paul, the agent in M-Ecolab [e.g., 32], is that neither Paul nor Scooter were designed explicitly to disrupt vicious cycles or to promote and create virtuous cycles. Both Paul and Scooter were designed with other goals in mind – in Scooter's case, addressing gaming the system, and in Paul's case encouraging behavior congruent with students' goal orientations.

Hence, it is possible that agents designed specifically to adapt to differences in affect (e.g., [4], [43]) may impact student affect to a greater degree than agents designed with other pedagogical goals foremost. It would be valuable to replicate the study conducted here with an agent designed specifically in this fashion -- if and when such a study is conducted, that study can be compared to the study presented here, in order to see whether different types of agents impact affect differently. In general, the issue of which affective responses and formative feedback by agents most impact student affect remains an open question. For example, several affective responses and forms of feedback might be appropriate responses to gaming the system (and are manifested by teachers when students game the system). Several automated responses to gaming have already been realized by different research groups (e.g., [58], [59], [60]), although these interventions have not yet been formally compared to each other. It will be a valuable area of future work to explicitly compare these different interventions, towards determining which forms of affective response and feedback may most impact the student affect underlying a behavior like gaming, and best address the behavior of gaming.

One key finding of this work (and prior research in affective dynamics (cf. [22], [24], [27]) is that student affect within learning software is quite stable, regardless of whether or not software agents are present; the same student often is experiencing the same affect in successive observations (180 seconds apart). In addition, while learning environments on different topics or used by different populations have very different base rates of different affective states, the same vicious cycles and virtuous cycles (particularly boredom and engaged concentration) are seen in many studies, involving a wide variety of types of educational software, from intelligent tutors to educational games. This commonality in findings suggests that there may be some universals in affect during learning, with affect being quite stable over time, regardless of the learning environment. Because of this, the first educational intervention that is concretely shown to radically alter students' moment-to-moment affect during learning, in particular by preventing vicious cycles and creating and reinforcing virtuous cycles, will have made a major contribution, and a major difference to learners.

Acknowledgment

This research project was supported by the Philippines Department of Science and Technology (DOST) Engineering Research and Development for Technology Consortium under the project "Multidimensional Analysis of User-Machine Interactions Towards the Development of Models of Affect", the DOST Philippine Council for Advanced Science and Technology Research and Development grant entitled "Development of Affect-Sensitive Interfaces", and the Pittsburgh Science of Learning Center, funded by the National Science Foundation, award #SBE-0836012. We thank Mrs. Carmela Oracion, Ivan Jacob Pesigan, Ramon Rodriguez, Thor Collin Andallaza, the Ateneo Center for Educational Development, the Department of Information Systems and Computer Science of the Ateneo de Manila University and the faculty, staff, and students of Ramon Magsaysay Cubao High School for their support in this project. We thank Lisa Rossi and Sidney D'Mello for helpful comments and suggestions.

References

[1] J. Cassell, J. Sullivan, S. Prevost, and E.F. Churchhill, *Embodied Conversational Agents*. Cambridge, MA: The MIT Press. 2000.

[2] C. Conati and X. Zhao, "Building and Evaluating an Intelligent Pedagogical Agent to Improve the Effectiveness of an Educational Game," in *9th International Conference on Intelligent User Interface*, , pp. 6-13, 2004.

[3] 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," in *Proc. 8th International Conference on Intelligent Tutoring Systems*, 2006, pp. 392 – 401.

[4] B. Woolf, W. Burleson, I. Arroyo, T. Dragon, D. Cooper and R. Picard, "Affect-aware tutors: recognizing andresponding to student affect," *International Journal of Learning Technology*, vol. 4, no. 3/4,pp. 129-163, 2009.

[5] J. Rickel and W. L. Johnson, “Integrating Pedagogical Capabilities in a Virtual Environment Agent.” *In Proc. of the First international Conf. on Autonomous Agents (Marina del Rey, California, United States, February 05 - 08, 1997).* AGENTS '97. ACM, New York, NY, 1997, pp. 30-38.

[6] W.Burleson, “Affective Larning Companions: Strategies for Empathetic Agents with Real-time Multimodal Affective Sensing to Foster Meta-cognitive and Meta-affective Approaches to Learning, Motivation, and Perseverance,”Ph. D. Dissertation. Massachusetts Institute of Technology, 2006.

[7] K. Leelawong and G. Biswas, “Designing Learning by Teaching Agents: The Betty’s Brain System,” *International Journal of Artificial Intelligence in Education*, vol. 18, no. 2, pp. 181-208, 2008.

[8] S. McQuiggan, J. Robison and J. Lester, “Affective Transitions in Narrative-centered Learning Environments”, in *Proc. of the 9th International Conf. on Intelligent Tutoring Systems,* 2008.

[9] M. M. T. Rodrigo, G. Rebolledo-Mendez, R. S. J. d. Baker, B. du Boulay, J. O. Sugay, S. A. L. Lim, M. B. E. Lahoz, R. Luckin, “The Effects of Motivational Modeling on Affect in an Intelligent Tutoring System,” *In Chan, T.-W., Biswas, G., Chen, F.-C., Chen, S., Chou, C., Jacobson, M., Kinshuk, Klett, F., Looi, C.-K., Mitrovic, T., Mizoguchi, R., Nakabayashi, K., Reimann, P., Suthers, D., Yang, S., and Yang, J.-C. (Eds.).International Conf. on Computers in Education*, 2008, pp. 49-56.

[10] B. Reeves, and C. Nass, *The Media Equation: How People Treat Computers, Television and New Media Like Real People and Places*, Center for the Study of Language and Information Publication Lecture Notes. Stanford: CA: CSLI Publications. 2001.

[11] N. Shechtman and L. M. Horowitz, , ”Media Inequality in Conversation: How People Behave Differently When Interacting with Computers and People.” *Proc. SIGCHI Conference on Human Factors in Computing Systems*, 2003, pp. 281-288.

[12] M.H. Bond, and C.K. Venus, “Resistance to Group or Personal Insults in an Igroup or Outgroup Context,” *International Journal of Psychology*, vol. 26, no. 1, pp. 83-94, 1991.

[13] C. Conati, and H. McLaren, “Empirically Building and Evaluating a Probabilistic Model of User Affect,” *User Modeling and User-Adapted Interaction*, vol. 19, no. 3, pp. 267-303, 2009.

[14] A. C. Graesser, S. K. D’Mello, S. D. Craig, A. Witherspoon, J. Sullins, B. McDaniel, and B. Gholson, “The relationship between affective states and dialog patterns during interactions with AutoTutor,” *Journal of Interactive Learning Research,* vol. 19, no. 2, 2008, pp. 293 – 312.

[15] G. Rebolledo-Mendez, B. du Boulay, and R. Luckin, “Motivating the Learner: An Empirical Evaluation,”*8th International Conference on Intelligent Tutoring Systems*, 2006, pp. 545-554.

[16] V. J. Shute, “Focus on Formative Feedback.” *Review of Educational Research*, vol. 78, no. 1 pp. 153-189, 2008.

[17] J. Cassell, and K.R. Thorisson, “The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents,” *Applied Artificial Intelligence*, vol. 13, pp. 519-538, 1999.

[18] C. Conati and M. Manske, “Evaluating Adaptive Feedback in an Educational Computer Game,” Proceedings of IVA 2009, *9th International Conf. on Intelligent Virtual Agents, Lecture Notes in Artificial Intelligence 5773*. Springer Verlag, 2009, pp. 146-158.

[19] S. Hidi, and V. Anderson, “Situational Interest and its Impact on Reading and Expository Writing,” in K. A. Renninger, S. Hidi, and A. Krapp, Eds., “*The role of interest in learning and development,”*Hillsdale, NJ: Erlbaum, 1992, pp. 215-238.

[20] A. Krapp, “Structural and Dynamic Aspects of Interest Development: Theoretical Considerations from an Ontogenetic Perspective,” *Learning and Instruction*, vol. 12, no. 4, pp. 383-409, 2002.

[21] R. Pekrun, T. Goetz, L, M. Daniels, R. H. Stupnisky, and R. P. Perry, “Boredom in Achievement Settings: Exploring Control–Value Antecedents and Performance Outcomes of a Neglected Emotion,” *Journal of Educational Psychology,* vol. 102, no. 3, pp. 531-549, 2010.

[22] R.W. Larson and M.H. Richards, “Boredom in the Middle School Years: Blaming Schools versus Blaming Students,” *American Journal of Education,* vol. 99, no. 4, pp. 418-443, 1991.

[23] D.W. Perkins, C. Hancock, R. Hobbs, F. Martin and R. Simmons, “Conditions of Learning in Novice Programmers,” Educational Technology Center, Office of Educational Research and Improvement, 1985.

[24] R. S. J. d. Baker, S. K. D’Mello, M. M. T. Rodrigo, A. C. Graesser, “Better to be Frustrated than Bored: The Incidence, Persistence, and Impact of Learners’ Cognitive-Affective States During Interactions with Three Different Computer-based Learning Environments,” *International Journal of Human-Computer Studies*, vol. 68, no. 4, pp. 223-241, 2010.

[25] S. K. D’Mello and A. C. Graesser, “Dynamics of Cognitive-Affective States During Deep Learning,” under review.

[26] S. K. D’Mello, R. Taylor, A. C. Graesser, “Monitoring Affective Trajectories During Complex Learning,” 29th *Annual Cognitive Science Soc.*, 2007, pp. 203 – 208..

[27] B. Kort, R. Reilly, and R. W. Picard, “An Affective Model of Interplay Between Emotions and Learning: Reengineering Educational Pedagogy - Building a Learning Companion,” In *International Conf. on Advanced Learning Technologies.* 2001.

[28] R. S. J. d. Baker, A. T. Corbett, K. R. Koedinger, and A. Z. Wagner, “Off-task Behavior in the Cognitive Tutor Classroom: When Students ‘Game the system’,” in *Proc. of ACMCHI2004: Computer-Human Interaction*, 2004, pp. 383-390.

[29] M. Cocea, A. Hershkovitz , and R.S.J.d Baker, “The Impact of Off-task and Gaming Behaviors on Learning: Immediate or Aggregate?” *Proc. 14th International Conf. on Artificial Intelligence in Education*, 2009, pp.507-514.

[30] 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, L. R. S. Sevilla, J. O. Sugay, S. Tep and N. J. B. Viehland, “Affect and Usage Choices in Simulation Problem-solving Environments,” In *R. Luckin, K.R. Koedinger, J. Greer(Eds),*

13th International Conf. on Artificial Intelligence in Education, pp. 145 – 152.

[31] R.S. Baker, "Designing Intelligent Tutors That Adapt to When Students Game the System," Ph.D. dissertation, CMU Technical Report CMU-HCII-05-104, 2005.

[32] K. R. Koedinger, "Toward Evidence for Instructional design Principles: Examples from Cognitive Tutor Math 6,"*Proc. of PME-NA XXXIII* (the North American Chapter of the Inter-national Group for the Psychology of Mathematics Education), Athens, GA, 2002, pp. 21-49.

[33] R. S. J. d. Baker, A. T. Corbett, I. Roll, K. R. Koedinger, "Developing a Generalizable Detector of When Students Game the System," *in User Modeling and User-Adapted Interaction: the Journal of Personalization Research*, vol. 18, no. 3, pp. 287-314, 2008.

[34] R. Inglehart, (2009, February 17), *Inglehart-Welzel Cultural Map of the World*. [Online]. Available:http://margaux.grandvinum.se/SebTest/wvs/articles/folder_published/article_base_54.

[35] C. Welzel , (2009, February 17), *A Human Development View on Value Change Trends [PowerPoint Presentation]*. Available:http://margaux.grandvinum.se/SebTest/wvs/articles/folder_published/article_base_83.

[36] Y. Uchida, V. Norasakkunkit, and S. Kitayama. "Cultural Constructions of Happiness: Theory and Empirical Evidence, *Journal of Happiness Studies*, vol *5*, pp. 223-239, 2004.

[37] F. Lynch. Social Acceptance Reconsidered. In A. A. Yengoyan and P. Q. Makil, Eds., *Philippine Society and the Individual: Selected Essays of Frank Lynch (1949-1976)*, USA: Center for South and Southeast Asian Studies, University of Michigan. 1984.

[38] C. Geertz, "'From the Native's Point of View': On the Nature of Anthropological Understanding," *Bulletin of the American Academy of Arts and Sciences*, 1974*, vol.* 28, no. 1, pp. 26-45.

[39] Ateneo Center for Educational Development. "Ramon Magsaysay Cubao High School: School Profile Report," Available from the Ateneo Center for Educational Development, Ateneo de Manila University, Loyola Heights, Quezon City, Philippines. 2009.

[40] T.W. Bickmore and R.W. Picard. "Towards Caring Machines," *CHI Extended Abstracts*, 2004, pp. 1489-1492.

[41] W.R. Cupach and B.H. Spitzberg, "Trait Versus State: A Comparison of Dispositional and Situational Measures of In-terpersonal Communication Competence," *The Western Journal of Speech Communication*, vol. 47, pp. 364-379, 1983.

[42] M. M. T. Rodrigo, R. S. J. d. Baker, S. D'Mello, M. C. T. Gonzalez, M. C. V. Lagud, S. A. L. Lim, A. F. Macapanpan, S. A. M. S. Pascua, J. Q. Santillano, J. O. Sugay, S. Tep, N. J. B. Viehland, "Comparing Learners' Affect While Using an Intelligent Tutoring System and a Simulation Problem Solving Game," *9th International Conf. on Intelligent Tutoring Systems*, pp. 40-49, 2008.

[43] S.K. D'Mello, R.W. Picard and A. Graesser, "Toward an Affect-Sensitive AutoTutor", *IEEE Intell. Syst.*, vol. 22, no. 4, pp. 53-61, 2007.

[44] S.K. D'Mello, S.D. Craig, A.W. Witherspoon, B.T. McDaniel, and A.C. Graesser, "Automatic Detection of Learner's Affect from Conversational Cues," *User Modeling and User-Adapted Interaction,* vol. 18, no. 1-2, pp. 45-80, 2008.

[45] A. De Vicente and H. Pain, "Informing the Detection of the Students' Motivational State: an Empirical Study,"*Proc. 6th International Conf. on Intelligent Tutoring Systems*, 2002, pp. 933-943.

[46] S.K. D'Mello and A.C. Graesser, "Automatic Detection of Learners' Emotions from Gross Body Language," *Applied Artificial Intelligence,* vol. 23, no. 2, pp. 123-150, 2009.

[47] S. Mota and R.W. Picard, "Automated Posture Analysis for Detecting Learner's Interest Level," *Proc. 2003 Conf. on Computer Vision and Pattern Recognition Workshop*, 2003.

[48] S.K. D'Mello, S.D. Craig, J. Sullins, and A.C. Graesser, "Predicting Affective States Expressed Through an Emote-Aloud Procedure from AutoTutor's Mixed-Initiative Dialogue," *International Journal of Artificial Intelligence in Educ.*, vol. 16, pp. 3-28, 2006.

[49] D.G. Cooper, I. Arroyo, B.P. Woolf, K. Muldner, W. Burleson, R. Christopherson, "Sensors Model Student Self-Concept in the Classroom," *Proc. International Conf. on User Modeling, Adaptation, and Personalization*, 2009, pp. 30-41.

[50] A.C. Graesser, B. McDaniel, P. Chipman, A. Witherspoon, S. D'Mello, and Gholson, B., "Detection of Emotions Duing Learning with AutoTutor," *Proc. 28th Annual Meeting of the Cognitive Science Society*, 2006, pp. 285-290.

[51] J. Schofield, *Computers and Classroom Culture.* Cambridge, MA: MIT Press. 1995

[52] H.M. Lahaderne, "Attitudinal and Intellectual Correlates of Attention: A Study of Four Sixth-Grade Classrooms," *Journal of Educational Psychology,* vol. 59, no. 5, pp. 320-324, 1968.

[53] N.L. Karweit and R.E. Slavin, "Time-On-Task: Issues of Time, Sampling, and Definition," *Journal of Experimental Psychology,* vol. 74, no. 6, pp. 844-851, 1982.

[54] S. Planalp, V.L. DeFrancisco, D. Rutherford, "Varieties of Cues to Emotion in Naturally Occurring Settings," *Cognition and Emotion,* vol. 10, no. 2, pp. 137-153, 1996.

[55] C.A. Bartel and R. Saavedra, "The Collective Construction of Work Group Moods," *Administrative Science Quarterly,* 2000, vol. 45, no. 2, pp. 197-231.

[56] B.A. Lehman, M. Mathews, S.K. D'Mello, and N. Person, "Understanding Students' Affective States During Learning," *Proc. 9th International Conf. on Intelligent Tutoring Systems*, 2008, pp. 50-59.

[57] M. Csikszentmihalyi, "*Flow: The Psyochology of Optimal Experience", *New York: Harper and Row, 1990.

[58] J. Walonoski and N. Heffernan, "Prevention of Off-task Gaming Behavior in Intelligent Tutoring Systems. In Ikeda, Ashley & Chan (Eds.). *Proc. 8th International Conf. on Intelligent Tutoring Systems*, pp. 722-724, 2006.

[59] I. Roll, V. Aleven, B.M. McLaren, and K.R. Koedinger, "Improving Students' Help-seeking Skills Using Metacognitive Feedback in an Intelligent Tutoring System," *Learning and Instruction*, vol. 21, pp. 267-280, 2011.

[60] I. Arroyo, K. Ferguson, J. Johns, T. Dragon, H. Meheranian, D. Fisher, A. Barto, S. Mahadevan, and B. P. Woolf,, "Repairing Disengagement with Non-Invasive Interventions," *Proc. 13th International Conf. of Artificial Intelligence in Edu-*

cation. IOS Press 2007.

[61] D'Mello, S. K., Taylor, R., Davidson, K., and Graesser, A. (2008). Self versus Teacher Judgments of Learner Emotions during a Tutoring Session with AutoTutor. *Proceedings of the Ninth International Conference on Intelligent Tutoring Systems,* pp. 9-18, 2008.

Ma. Mercedes T. Rodrigo is an Associate Professor at the Department of Information Systems and Computer Science and the head of the Ateneo Laboratory for the Learning Sciences. She received her PhD in Computer Technology in Education from Nova Southeastern University in 2002, her MS in Applied Computer Science from the University of Maryland Eastern Shore in 1992 and her BS in Computer Science (honorable mention) from the Ateneo de Manila University in 1988. From 2003-2008 was the Chair of the Department of Information Systems and Computer Science. In 2008, Dr. Rodrigo was a Fulbright Senior Research Fellow at the Pittsburgh Science of Learning Center. Dr. Rodrigo's areas of interest are affective computing and artificial intelligence in Education.

Ryan S. J. d. Baker is Assistant Professor of Psychology and the Learning Sciences at Worcester Polytechnic Institute. Previously he was Technical Director of the Pittsburgh Science of Learning Center DataShop. He received his Ph.D. and M.S. in Human-Computer Interaction in 2005 from Carnegie Mellon University and his Sc.B. in Computer Science in 2000 from Brown University. He is President of the International Educational Data Mining Society. His work on an agent that responded to gaming the system won the Best Paper Award at the 8th International Conference on Intelligent Tutoring Systems in 2006.

Jenilyn Agapito obtained her MS in Computer Science in 2011 from the Ateneo de Manila University. She received her BS in Computer Science from Ateneo de Naga University in 2007. She was Assistant Instrucor I at the Ateneo de Naga from 2007-2009. She has expressed her interest in the areas of affective computing and educational data mining..

Julieta Nabos is a graduate student at the Department of Information Systems and Computer Science in Ateneo de Manila University. She received her Master of Education (major in Management) from Marinduque State College in 2006 and BS in Computer Science from Eulogio "Amang" Rodriguez Institute of Science and Technology in 1994. She is an Instructor at the School of Information and Computing Sciences in Marinduque State College.

Ma. Concepcion Repalam is an instructor at Laguna State Polytechnic University. She obtained her Masters in Information Technology at the Ateneo de Manila University in 2010. Her interests are the use of technology in education and affective computing.

Salvador S. Reyes, Jr. is pursuing his Master's Degree in Computer Science at the Ateneo de Manila University. He graduated with a bachelor's degree in Computer Science from the same university in 2009.

Ma. Ofelia C. Z. San Pedro is pursuing a doctoral degree in the learning sciences and technologies at Worcester Polytechnic Institute. She is also a Research Assistant and the WPI Educational Psychology Laboratory. She obtained her MS degree in Computer Science at the Ateneo de Manila University in 2011. She received a BS degree in Electronics and Communications Engineering from the same university 2005 Her research interests include data mining, machine learning, intelligent tutoring systems, human-computer interaction and software engineering. She previously worked in the area of software engineering providing software and usability solutions to improve different business processes. Her current research work is focused on exploring the model of carelessness as a student behavior with the use of an intelligent tutoring system**.**