Int J Artif Intell Educ (2014) 24:189–210
DOI 10.1007/s40593-014-0015-y

RESEARCH ARTICLE

Carelessness and Affect in an Intelligent Tutoring System for Mathematics

Maria Ofelia Z. San Pedro · Ryan S. J. d. Baker · Ma. Mercedes T. Rodrigo

Published online: 26 February 2014
© International Artificial Intelligence in Education Society 2014

Abstract We investigate the relationship between students' affect and their frequency of careless errors while using an Intelligent Tutoring System for middle school mathematics. A student is said to have committed a careless error when the student's answer is wrong despite knowing the skill required to provide the correct answer. We operationalize the probability that an error is careless through the use of an automated detector, developed using educational data mining, which infers the probability that an error involves carelessness rather than not knowing the relevant skill. This detector is then applied to log data produced by high-school students in the Philippines using a Cognitive Tutor for scatterplots. We study the relationship between carelessness and affect, triangulating between the detector of carelessness and field observations of affect. Surprisingly, we find that carelessness is common among students who frequently experience engaged concentration. This finding implies that a highly engaged student may paradoxically become overconfident or impulsive, leading to more careless errors. In contrast, students displaying confusion or boredom make fewer careless errors. Further analysis over time suggests that confused and bored students have lower learning overall. Thus, their mistakes appear to stem from a genuine lack of knowledge rather than carelessness.

Keywords Carelessness · Affect · Cognitive tutor · Educational data mining

M. O. Z. San Pedro (✉) · R. S. J. d. Baker
Department of Human Development, Teachers College, Columbia University, 525 W 120th Street, New York, NY 10027, USA
e-mail: mzs2106@tc.columbia.edu

R. S. J. d. Baker
e-mail: baker2@exchange.tc.columbia.edu

M. M. T. Rodrigo
Department of Information Systems and Computer Science, Ateneo de Manila University, Katipunan Ave, Loyola Heights, Quezon City 1108, Philippines
e-mail: mrodrigo@ateneo.edu

Introduction

Disengagement among students has been shown and continues to be a major problem in education. Disengagement, manifested in various fashions, can reduce learning outcomes (Baker et al. 2004; Aleven et al. 2006; Balfanz et al. 2007; Lau and Darmanegara 2007; Cocea et al. 2009; Rowe et al. 2009), and at its extreme can lead to failure in courses, school drop-out, and failure to attend university (Rumberger and Larson 1998; Balfanz et al. 2007; San Pedro et al. 2013). As such, reducing student disengagement has become a goal for policy-makers (cf. National Research Council and Institute of Medicine 2004). In order to accomplish this goal, there has been increasing attention to the factors underlying the development of student disengagement (cf. Nottelmann and Hill 1977; Ryan and Patrick 2001; Balfanz et al. 2007; Baker et al. 2008c).

Though improved engagement has been seen as part of the promise of computer-aided instruction (Becker 2000; Sosa et al. 2011), disengagement appears to be a problem among students using various forms of computer-aided instruction, even when the instruction as a whole is seen as quite engaging (cf. Schofield 1995). Students have been reported to disengage in several ways while using computer-aided instruction, including 1) "gaming the system", exploiting properties of the system to get the solution, rather than by learning, which includes behaviors such as systematic guessing (Baker et al. 2004) and misuse of software help features (e.g. Aleven et al. 2004), and 2) engaging in off-task behavior (Karweit and Slavin 1981), i.e., surfing the web, talking to a friend about something other than the learning material, 3) working carelessly, making errors that do not reflect the student's knowledge (Hershkovitz et al. 2011), and 4) using learning systems in a way unrelated to the educational or stated task (Rowe et al. 2009; Wixon et al. 2012), variously termed off-task behavior or "without thinking fastidiously" behavior.

There has been evidence in recent years that these disengaged behaviors appear to emerge following the display of negative affect by students. For instance, frustration has been found to be associated with gaming the system (Baker et al. 2008c) and boredom has been found to precede gaming the system (Baker et al. 2010b). Additionally, boredom (Baker et al. 2011) has been found to precede off-task behavior, and confusion (Sabourin et al. 2011) has been found to precede using the learning system in a way unrelated to the educational task. These behavior-affect dynamics show that affect and disengagement interact very closely. However, it is not yet known how affect interacts with carelessness.

Carelessness

Carelessness, despite its perceived importance among many educators since the 1950s (cf. Eaton et al. 1956), and recent evidence that it is associated with the failure to attend college (San Pedro et al. 2013), has been relatively lightly studied in education. There have been three paradigms for studying carelessness among learners: analyzing error patterns using heuristics, self-report measures, and analyzing error patterns using data mining.

The first paradigm, analyzing error patterns within paper mathematics tests using heuristics, emerges from work by Clements in the 1980s. Clements (1982) studied 6th-grade students' errors during mathematics problem-solving. Clements administered

mathematical assessment tests twice to 50 sixth-grade students on successive days, and items that the student answered correctly on one occasion but incorrectly on the other. They were then interviewed, using a structured method for analyzing problem-solving process during interviews (Newman 1977). If the student got an incorrect answer during the interview as well, then the original error was deemed non-careless. But if the student got the problem correct during the interview without any assistance from the interviewer, then the original error was deemed careless. Through this method, Clements assessed 20 % of student errors as careless. Interestingly, Clements found that over-confidence was associated with carelessness.

The second paradigm, using self-report measures, was developed by Maydeu-Olivares and D'Zurilla (1996), who developed a scale to represent students' perceptions that their problem-solving attempts are impulsive, careless and incomplete. Their self-report measure of carelessness was found to be correlated with academic performance, as it was shown to be an effective predictor of grades.

A third paradigm is to analyze error patterns using data mining. In this paradigm, an automated detector of carelessness is developed, which attempts to assess whether an error is due to a lack of knowledge or carelessness. This model is conceptually similar to Clements' approach, defining careless errors as errors where the student knows how to answer correctly, but is based on data mining rather than heuristics. The detection approach will be discussed later, but relies upon assessments of the probability that the student knows the relevant skill (cf. Corbett and Anderson 1995), and data on future performance in the tutoring software (Baker et al. 2008b). This detector has been shown to predict post-test performance, even after taking assessments of student skill into account (Baker et al. 2010a).

Beyond educational settings, careless errors have been studied in the literature on slips, defined by Norman (1981) as an error where a person makes an action that is not intended. Slips are seen as involving inattention, or even "absent-mindedness" in some theoretical accounts (Eysenck and Keane 1990). Hay and Jacoby (1996) stated that slips are most likely to happen when the correct response was not the strongest, and when the response had to be acted on rapidly where attention is not most likely to be directed to the correct response. By contrast, Norman (1981) theorized that action slips occur when there are errors in intention formation, activation of the wrong schema, or actions determined by a wrong schema. Norman theorized that action slips may be caught by the person just before they make an error, at the moment the error occurs. But they also may never be caught at all (Norman 1981). The factors leading to carelessness in educational settings are not yet fully known, though there has been some research on the relationship between individual differences and carelessness. Maydeu-Olivares and D'Zurilla (1996) conceptualize carelessness as a stable personality trait, and Hershkovitz et al. (2011) find evidence linking carelessness to having either learning goals or performance goals, with students manifesting neither goal unexpectedly demonstrating more careful behavior during learning. Clements (1982) finds links between over-confidence and carelessness. Baker and Gowda (2010) find year-long differences in the prevalence of carelessness between urban, rural, and suburban students. At the same time, even if there are stable individual differences in carelessness, no student is always careless. Hence, there must be triggering factors leading to students becoming careless in specific situations. One such factor potentially leading to carelessness is student affect, given that emotions may also influence an individual's

decision making and learning (Adolphs and Damasio 2001). As discussed earlier, other forms of disengaged behavior such as off-task behavior and gaming the system have strong links to student affect. It is reasonable to hypothesize similar relationships for carelessness. For instance, Epstein (1979) hypothesizes that careless behaviors among learners may emerge due to feelings of confusion and tension.

Within this paper, we study this issue within the context of students using an AIED learning environment. AI-based learning environments are becoming increasingly prominent within education (cf. Mitrovic 2003; Suraweera and Mitrovic 2004; Razzaq et al. 2005; Koedinger and Corbett 2006; Heffernan et al. 2008; Mitrovic et al. 2009; Razzaq and Heffernan 2009; Sosa et al. 2011), and many such environments provide detailed log files which can support fine-grained analysis through educational data mining techniques (Baker and Yacef 2009). More specifically, we will study these issues within a Cognitive Tutor (Koedinger and Corbett 2006). Cognitive Tutors provide students with guided learning support as they engage in problem-solving, and provide excellent-quality logs that have formed the basis of dozens of data mining studies (Koedinger et al. 2010). In addition, models have already been developed for Cognitive Tutors that can assess the probability that an error is not due to student knowledge (cf. Baker et al. 2008b), based on models of student learning that can estimate the knowledge state of each student at a given time (Corbett and Anderson 1995), considerably facilitating research into the factors promoting carelessness.

In this paper, we use a detector of carelessness to study the relationship between affect and carelessness. We do so using data from a population of students in the Philippines, for whom field observations of affect are available. To use the carelessness detector within this population, we validate that the detector can accurately predict student errors when applied to students from a different country, studying generalizability between the USA and Philippines. We then apply the carelessness detector to the log files from this data set, labeling every student action in terms of carelessness, in order to assess each student's degree of carelessness. We then assess which affective states are associated with careless errors through correlational analysis. Finally, we examine whether the relationships between affective states and careless errors change over time. This builds on our previous research (San Pedro et al. 2011a; San Pedro et al. 2011b) that started to investigate carelessness using a Cognitive Tutor unit. In this paper, we present significant enhancements, substantial discussion on this behavior within a learning environment, and a more comprehensive theoretical perspective about carelessness.

Method

Data were gathered from 126 public high school students in Quezon City, Philippines (labeled PH below), who used a Cognitive Tutor unit on scatterplot generation and interpretation (Baker et al. 2006) for 80 min, inside the school's computer center, after school. The students were between 12 and 14 years old. Students had not explicitly covered these topics in class prior to the study. Prior to using the software, students viewed conceptual instruction, delivered via a PowerPoint presentation with voiceover and some simple animations. Each student in each class took a nearly isomorphic pre-test and post-test, counterbalanced across conditions.

Within the Scatterplot Tutor (Figs. 1, 2, and 3), the learner is given a problem scenario. He/she is also provided with data that he/she needs to plot in order to arrive at the solution. He/she is asked to identify the variables that each axis will represent. He/she must then provide an appropriate scale for each axis. He/she has to label the values of each variable along the axis and plot each of the points of the data set. Finally, he/she interprets the resultant graphs. The Scatterplot tutor provides contextual hints to guide the learner, feedback on correctness, and messages for errors. The skills of the learner are monitored and displayed through skill bars that depict his/her mastery of skills.

Sixty four of the participants (referred to as the Experimental or Scooter group) were randomly assigned to use a version of the tutor with a pedagogical agent, "Scooter the Tutor", shown in Fig. 4. Scooter was designed to 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. Previous research in the United States found that Scooter reduced gaming, while increasing learning for gaming students (Baker et al. 2006); however, these patterns were not statistically significant in the Philippines (Rodrigo et al. 2012). The remaining 62 participants (Control or NoScooter group) used a version of the Scatterplot Tutor without the pedagogical agent. The number of students assigned to the conditions in this study was unbalanced because of data gathering schedule disruptions caused by inclement weather.

Data on student carelessness were produced from the logs generated from Cognitive Tutor usage, while data on affective states were collected using the BROMP protocol for quantitative field observation (cf. Ocumpaugh et al. 2012). Student affective state was coded by a pair of expert field observers as students used the tutor. Each observation lasted up to twenty seconds, with each participant being observed 24 times having an interval of 180 s between observations. Each observation was conducted using side glances, to reduce observer effects. To increase tractability of both coding and eventual analysis, if two distinct affective states were seen during a single observation, only the first state observed was coded. Any affective state of a student other than the student currently being observed was not coded. The observers based their judgment of a student's affective state on the student's work context, actions, utterances, facial expressions, body language, and interactions with teachers or fellow

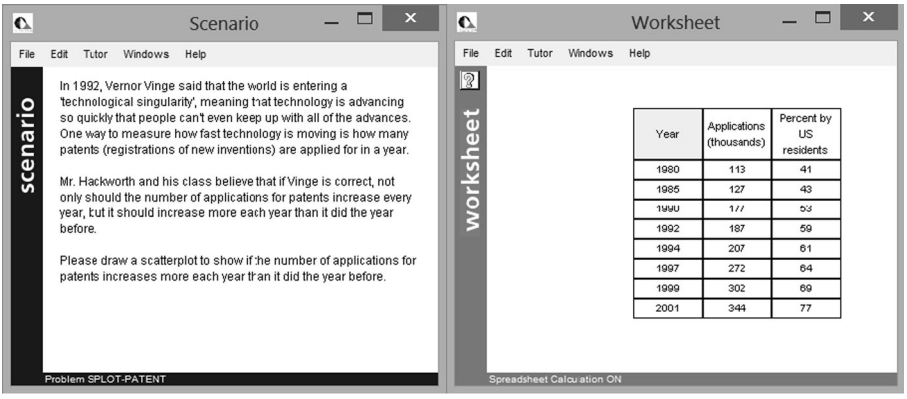

Fig. 1 Scatterplot problem scenario (**a**) and data set (**b**)

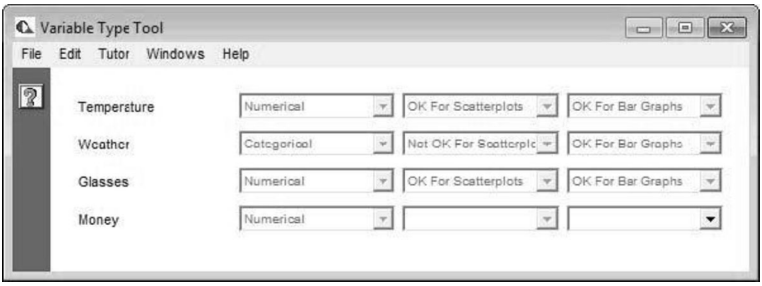

Fig. 2 Scatterplot variable type tool

students. These are, broadly, the same types of information used in previous methods for coding affect (e.g., Bartel and Saavedra 2000), and in line with Planalp et al.'s (1996) descriptive research on how humans generally identify affect using multiple cues in concert for maximum accuracy rather than attempting to select individual cues. The observers' inter-rater reliability was found to be moderate, with Cohen's (1960) $\kappa=0.54$.

The coding scheme included seven categories: boredom (Pekrun et al. 2010), confusion (Kort et al. 2001; Craig et al. 2004), delight (Fredrickson and Branigan 2005), engaged concentration (the affective state associated with Csikszentmihalyi 1990's construct of "flow"), frustration (Kort et al. 2001), surprise (Schutzwohl and Borgstedt 2005), and a category defined as '?' representing any affect other than the previous six mentioned. These categories are referred to as cognitive-affective states (Baker et al. 2010b), as they are states found to involve both cognitive and affective aspects. These affective states were chosen due to arguments that they are more representative of affect during learning, than Ekman and Friesen's (1978) six basic emotions of fear, anger, happiness, sadness, disgust and surprise (Csikszentmihalyi 1990; Kort et al. 2001; Craig et al. 2004; D'Mello et al. 2010; D'Mello and Graesser

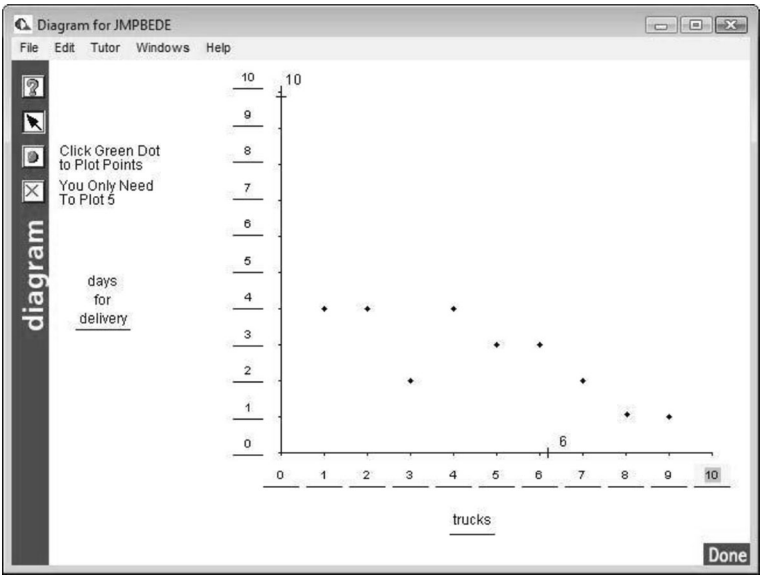

Fig. 3 Scatterplot user interface for point-plotting

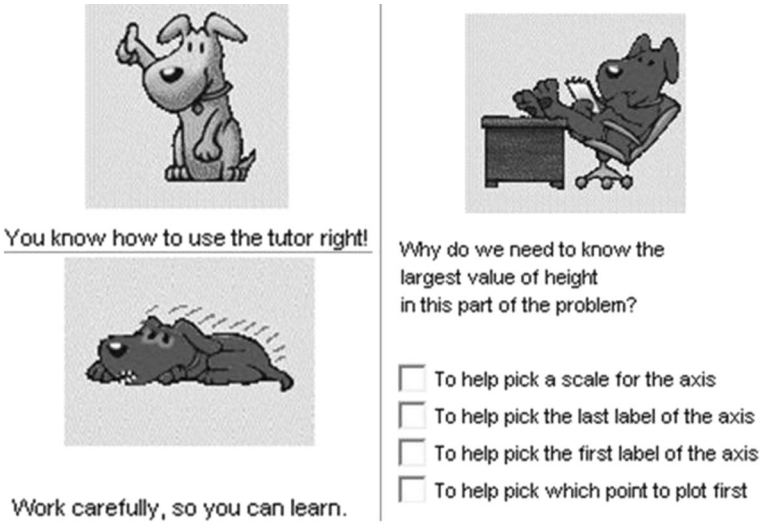

Fig. 4 Scooter the tutor's changes in behavior – *top left*: scooter is pleased with the student's appropriate use of the tutor software; *bottom left*: scooter is angry with heavy gaming behavior; *Right*: Scooter offers a supplementary exercise to a gaming student

2012). Such affective states are also persistent when solving problems within computer learning environments (Arroyo et al. 2009; D'Mello et al. 2006, 2007, 2010; Graesser et al. 2006; Woolf et al. 2009). Engaged concentration is in specific quite common during learning (cf. Baker et al. 2010b), and refers to the state of engagement with a task where concentration is intense, attention is focused, and involvement is complete (Baker et al. 2010b), while not including task-related aspects of Csikszentmihalyi's (1990) construct of flow such as clear goals, balanced challenge, and direct and immediate feedback.

Despite being learning-centered, this set of cognitive-affective states can also be mapped within the perspective of emotions using Russell's (2003) Core Affect framework, as discussed in (Baker et al. 2010b). Russell's Core Affect framework asserts that an affective state has two components: valence (increasing or decreasing feelings of pleasure) and arousal (increasing or decreasing feelings of activation and energy). Baker and colleagues hypothesized that boredom has negative valence and low arousal; confusion has negative valence and moderate arousal; frustration has high negative valence and a high arousal; delight has positive valence and high arousal; surprise has high arousal and can have either positive or negative valence; and engaged concentration has positive valence. It is not yet clear what the level of arousal is for engaged concentration, as engaged concentration can be an intense experience, but there can also be a decrease in arousal when an individual experiences uninterrupted and concentrated cognition (cf. Mandler 1984).

Operationalization of Carelessness Using Logs

The incidence of carelessness within the Cognitive Tutor was traced with a model designed to assess "slips" (cf. Baker et al. 2008b). Slips in that paper are operationalized in a fashion essentially identical to prior theory of how to identify careless errors

(Clements 1982). The model used in (Baker et al. 2008b), termed the Contextual Slip model, contextually estimates the probability that a specific student action indicates a slip/carelessness, whenever the student reaches a problem step requiring a specific skill, but answers incorrectly. The probability of carelessness/slip is assessed contextually, and is different depending on the context of the student error. The probability estimate varies based on several features of the student action and the situation in which it occurs, including the speed of the action, and the student's history of help-seeking from the tutor. As such, the estimate of probability of carelessness/slip is different for each student action. The Contextual Slip model has been shown to be a statistically significant predictor of student performance on a post-test measuring learning from a Cognitive Tutor for Genetics, even after controlling for assessment of each student's knowledge within the software (Baker et al. 2010a).

The Contextual Slip model is created using Bayesian Knowledge Tracing (BKT) (Corbett and Anderson 1995), a model used to estimate a student's latent knowledge based on his/her observable performance. The BKT framework, in its original articulation, is used within Cognitive Tutors to infer student knowledge by continually updating the estimated probability a student knows a skill every time the student gives a first response to a problem step. It uses four parameters – two learning parameters L_0 (initial probability of knowing each skill) and T (probability of learning the skill at each opportunity to make use of a skill), together with two performance parameters G (probability that the student will give a correct answer despite not knowing a skill) and S (probability that the student will give an incorrect answer despite knowing the skill). Each of the four parameters are estimated separately for each skill, thus each skill has its own BKT model (estimated from data for each skill). Note that in this model – different from the Contextual Slip model, though BKT is used to create that model – the four parameters for each skill are invariant across the entire context of using the tutor, and invariant across students. Using Bayesian analysis, BKT re-calculates the probability that the student knew the skill before the response (at time n-1), using the information from the response, then accounts for the possibility that the student learned the skill during the problem step, such that (Corbett and Anderson 1995):

$$P\left(L_{n-1}\middle|Correct_n\right) = \frac{P(L_{n-1})*(1-P(S))}{P(L_{n-1})*(1-P(S)) + (1-P(L_{n-1}))*P(G)}. \quad (1)$$

$$P\left(L_{n-1}\middle|Incorrect_n\right) = \frac{P(L_{n-1})*P(S)}{P(L_{n-1})*P(S) + (1-P(L_{n-1}))*(1-P(G))}. \quad (2)$$

$$P\left(L_n\middle|Action_n\right) = P\left(L_{n-1}\middle|Action_n\right) + \left(\left(1-P\left(L_{n-1}\middle|Action_n\right)\right)*P(T)\right). \quad (3)$$

With the information from the logs, parameters needed for a baseline BKT model were fitted by employing brute-force grid search (cf. Baker et al. 2010a), an approach that finds optimal parameter estimates for problems where local optimization methods

such as iterative gradient descent are likely to produce relatively poor estimates. From this baseline model, we then create first-step estimations of the probability that each action is a contextual slip. These estimations are not the final Contextual Slip model, but are used to produce it. Specifically, we use BKT to estimate whether the student knew the skill at each step. In turn, we use these estimates, in combination with Bayesian equations, to label incorrect actions with the probability that the actions were slips, based on the student performance on successive opportunities to apply the rule. More specifically, given the probability that the student knows the skill at a specific time, Bayesian equations and the static BKT parameters are utilized to compute labels for the Slip probabilities for each student action (A) at time N, using future information (two actions afterwards – N+1, N+2). In this approach, we infer the probability that a student's incorrectness at time N was due to not knowing the skill, or whether it is due to a slip. The probability that the student knew the skill at time N can be calculated, given information about the actions at time N+1 and N+2 ($A_{N+1,N+2}$), and the other parameters of the Bayesian Knowledge Tracing model:

$$P\left(A_N \text{ is a Slip} \middle| A_N \text{ is incorrect}\right) = P\left(L_n \middle| A_{N+1,N+2}\right). \quad (4)$$

This gives us a first estimate that a specific incorrect answer is a slip. However, this estimate uses data on the future, making it impossible to use to assess slip in real-time. In addition, there is considerable noise in these estimates, with estimates trending to extreme values that over-estimate slip in key situations due to limitations in the original BKT model (Baker et al. 2008a). But these estimated probabilities of slip can be used to produce a less noisy model that can be used in real time, by using them as training labels (e.g. inputs) to machine-learning. Specifically, a linear regression model is created that predicts slip/carelessness contextually. The result is a model that can now predict at each practice opportunity whether an action is a slip, using only data about the action itself, without any future information.

Models for Contextual Slip were produced using truncated training data (cf. Baker et al. 2008a), to predict contextual slip without using data from the future. This adjustment of the training data improves prediction of future performance by removing training examples where the prior probability, $P(L_{n-1})$, that the student knew the skill was extremely high, approaching 1. When $P(L_{n-1})$ approaches 1, there can be over-estimation of the probability that an error is due to a slip, if the estimate of student knowledge is erroneously high.

In order to validate these detectors' ability to predict future student correctness (a validation of their predictive accuracy), we substitute the contextual slip model, and a corresponding contextual guess model (cf. Baker et al. 2008b) for the Slip and Guess parameters of Bayesian Knowledge Tracing, labeling each student action with variant estimates as to how likely the response is a guess or a slip.

Creation and Validation of Carelessness Detectors

Using the Contextual Slip approach detailed in the previous section, we created our carelessness detectors by training models with each student action originally labeled with the probability estimate of slip occurrence, using information on that student

action generated on our tutor logs. For each action, a set of numeric or binary features were distilled, describing that action. These features, drawn from previous work to detect contextual slips in other learning software (e.g. Baker et al. 2008b), included:

- Action assessment – correct, incorrect, known bug (procedural misconception), or a help request.
- The type of interface widget involved in the action – was the student selecting from a checkbox, pull-down menu, typing in a number or a string, or plotting a point.
- Was this the student's first attempt to act (answer or ask for help) on this problem step?
- The tutor's assessment of the probability that the student knows the skill involved in this action, called "p-know", derived using the Bayesian Knowledge Tracing algorithm in (Corbett and Anderson 1995).
- "Pknow-direct" - if the current action is the student's first attempt on this problem step, then pknow-direct is equal to p-know, but if the student has already made an attempt on the problem step, then pknow-direct is −1.
- Is this the first attempt of student to answer or get help on problem step?
- Time taken for the current action in seconds.
- The time taken for the action, expressed in terms of the number of standard deviations this action's time was faster or slower than the mean time taken by all students on this problem step, across problems.
- The time taken in the last 3 actions, expressed as the sum of the numbers of standard deviations each action's time was faster or slower than the mean time taken by all students on that problem step, across problems.
- Time taken by this student on each opportunity to practice the skill involved in the action, averaged across problems.
- Percentage of past problems where the student made errors on this problem step.
- The number of times the student has gotten this specific problem step wrong, across all problems.
- The number of times the student asked for help or made errors on this skill, across all problems.
- The number of times in the 3 previous actions where the action involved the same interface widget as the current action.
- The number of times in the 5 previous actions where the action involved the same interface widget as the current action. The number of times in the previous 8 actions where the action was a help request.
- The number of times in the previous 5 actions where the action was incorrect.

As in previous work to model slipping, the features extracted from each student action within the tutor were used to predict the probability that the action represents a slip. The prediction took the form of a linear regression model, fit using M5-prime feature selection in the RapidMiner 5.0 data mining package (an earlier version of this system is described in Mierswa et al. 2006). This resulted in numerical predictions of the probability that a student action was a careless error, each time a student made a first attempt on a new problem step. Linear regression was chosen as an appropriate modeling framework when both predictor variables and the predicted variable are numeric. In addition, linear regression is known to function well with noisy educational

data, creating relatively low risk of finding an "over-fit" model that does not function well on new data (cf. Hawkins 2004).

Six-fold student-level cross-validation (cf. Efron and Gong 1983) was conducted to evaluate the detector's goodness. Within this approach, a model is repeatedly trained on five groups of students and tested on a "held-out" sixth group of students. Cross-validating at this level allows us to assess whether the model will remain effective for new students drawn from the same overall population of students studied. Models were trained separately on the two groups of students using the two versions of the software. Table 1 shows a model trained on data that used the tutor without the pedagogical agent

Table 1 Carelessness (contextual slip) models for the noscooter and scooter groups

Carelessness (NoScooter group) =	Carelessness (Scooter group) =
−0.00942 * Action is a bug	+0.03783 * Action is a bug
+0.08783 * Action is a help request	+0.08464 * Action is a help request
+0.02517 * Input is a string	−0.05084 * Input is a string
−0.00500 * Input is a number	+0.01546 * Input is number
+0.01593 * Input is a point	+0.08198 * Input is a point
−0.02812 * Input is checkbox or not choice/string/ number/point	−0.08615 * Input is checkbox or not choice/string/ number/point
+0. 96077 * Probability that the student knew the skill involved in this action	+0.01126 * Probability that the student knew the skill involved in this action
−0.03118 * Not first attempt at skill in this problem	+ 0.01126 * "Pknow-direct"
+0.00019 * Time taken	−0.02912 * Not first attempt at skill in this problem
−0.01451 * Time taken, normalized in terms of SD off average across all students at this step	+0.00115 * Time taken
	−0.00820 * Time taken, normalized in terms of SD off average across all students at this step
+0.00612 * Time taken in last three actions, normalized	−0.00283 * Time taken in last three actions, normalized
−0.00268 * Time taken in last five actions, normalized	−0.00197 * Time taken in last five actions, normalized
+0.00072 * Number of errors the student made on this step on all problems	−0.00257 * Number of errors the student made on this step on all problems
+0.00511 * Percentage of past problems the student asked for help on this problem step	−0.00906 * Percentage of past problems the student asked for help on this problem step
+0.00021 * Percentage of past problems the student made errors on this problem step	−0.00121 * Percentage of past problems the student made errors on this problem step
−0.00004 * Time taken by this student on each opportunity for this problem step	−0.00099 * Time taken by this student on each opportunity for this problem step
−0.00734 * How many of the previous 5 actions were errors	−0.00999 * How many of the previous 8 actions were help requests
+0.02910 * Probability of knowing the skill before answering	+0.00354 * How many of the previous 5 actions were errors
−0.00366	+0.81721 * Probability of knowing the skill before answering
	+0.01842

(NoScooter group) and a model trained on data that used a tutor with an agent (Scooter group), with their respective final attributes. The detector from the NoScooter group data achieved a cross-validated correlation coefficient of r=0.886 to the original training labels of the probability that each student action was a slip, while the detector from the Scooter group data achieved r=0.836, in each case a high degree of correlation (Rosenthal and Rosnow 2008).

Generalizability of Carelessness Detectors

Aside from assessing a model's ability to predict existing data, model evaluation and selection should also include the ability of a model to generalize to different contexts and domains (Forster 2000; Myung et al. 2005). We have already assessed one form of generalizability for our models, by cross-validating them at the student level, showing predictive accuracy within new students from the student population, using the same tutor interface. In this section, we evaluate the generalizability of our models when applied to data from students using a different tutor interface, and to data from a different student population.

To investigate generalizability of our carelessness detectors to data from students using a different tutor interface, we tested each detector on the other data set, i.e. the NoScooter detector was tested on the Scooter group dataset and the detector from the Scooter group was tested on the NoScooter group dataset. Table 2 shows the detectors' correlation between the labeled (from Eq. 2) and predicted (from the models) values of slip, for each student action, in each data set. Within the NoScooter condition data, the detector trained on the Scooter condition data performed comparably or perhaps slightly lower (but still good, r=0.846) than the detector trained on the NoScooter data (r=0.886). Within the Scooter data, the detector trained on the NoScooter data performed worse (r=0.484) than the detector trained on the Scooter data (r= 0.836), although still respectably. These results appear to indicate that there is some degradation when a carelessness detector is transferred between versions of the tutor with or without a pedagogical agent, but that models remain substantially above chance. One possible interpretation for the asymmetry in transfer between the two environments is that the skills and problem steps in the NoScooter environment are also present in the Scooter environment, whereas the opposite is not true: Scooter provides additional supplementary exercises to students when they game the system. As such, the model trained from the Scooter group accounted for additional problem steps not present in the NoScooter group; the NoScooter model, not trained on these steps, could not predict gaming during these steps as well.

Table 2 Correlation (r value) of slip detectors to slip labels in different data sets

	NoScooter-group detector (PH)	Scooter-group detector (PH)
NoScooter Group Data (PH)	0.886	0.846
Scooter Group Data (PH)	0.484	0.836
NoScooter Group Data (US)	0.708	0.834
Scooter Group Data (US)	0.609	0.788

We also established the detectors' generalizability to new data from new student populations by applying our detectors trained on data from students in the Philippines to Scatterplot log data from a school in the USA. These interaction logs from the USA (described in greater detail in Baker et al. 2006) were gathered from 6th–8th grade students, in the suburbs of a medium-sized city in the Northeastern USA. Fifty-two students used the Scooter version of the tutor, and 65 students used the NoScooter version. The same features were distilled from the USA student logs from both the NoScooter and Scooter tutor versions, as had been distilled from the data from the Philippines. The detectors trained on data from the Philippines, were applied without modification to these distilled data from the USA students, to test for generalizability.

When transferred to data from the USA, both of the detectors trained on data from the Philippines performed quite well for all combinations of training and test conditions. The detector developed on the NoScooter sample from the Philippines achieved a correlation of 0.708 to the USA NoScooter data, and a correlation of 0.609 to the USA Scooter data. The detector developed on the Scooter sample from the Philippines achieved a correlation of 0.834 to the USA NoScooter data, and a correlation of 0.788 to the USA Scooter data. This is evidence for detector generalizability, as the detectors perform comparably well (when applied to the USA NoScooter data) or a little worse but still very respectably (when applied to the USA Scooter data), in a new country than in the original country, with no re-fitting. As a whole, taking correlation as a metric, the carelessness detectors trained in this study appear to show little degradation when transferred between these two countries. Models of help-seeking skill have also been shown to transfer between the USA and Philippines (Soriano et al. 2012), suggesting some commonalities in how behavior manifests between these two countries.

Studying the Relationship Between Carelessness and Affect

Having developed and validated a detector of student carelessness, we can now utilize it to analyze the relationship between carelessness and affect, in the data set from the Philippines for which affective observational data was available. The first step to doing this was to apply the detector trained for each condition (Scooter) and (NoScooter) to the data from that condition. For this analysis, we assessed the carelessness of each student in each group (e.g. Scooter and NoScooter), by taking the average probability of carelessness (slip estimates) on each incorrect action the student made, as in (Baker et al. 2010a; Baker and Gowda 2010).

There was an overall difference in carelessness between the two conditions. The overall mean carelessness (through slip probability estimates) for students in the NoScooter environment was 0.41. This does not mean that 41 % of errors were careless errors, but that students have a probability of 41 % of making a slip if they know the skill. The Scooter group had an overall mean carelessness of 0.36. The difference in carelessness between the two conditions was significant, $t(124)=2.01$, two-tailed $p=0.047$. This suggests that the pedagogical agent, designed to reduce gaming, has somehow reduced carelessness. At the same time, there were no significant differences in the frequency of any affective state between conditions. The largest difference found between conditions was for Engaged Concentration. Students in the Scooter condition

displayed this affect 37.17 % of the time; students in the NoScooter condition displayed this affect 43.45 % of the time; $t(124)=1.52$, two-tailed $p=0.13$.

We studied the relationship between each student's overall proportion of carelessness and their overall proportion of each affective state, with correlational analyses conducted in SSPS, shown in Table 3. The results were somewhat surprising. Carelessness was negatively correlated with boredom in both interfaces, $r=-0.36$, $p=0.004$ for NoScooter group; $r=-0.35$, $p=0.004$ for Scooter group. Confusion was also negatively correlated with carelessness, $r=-0.38$, $p=0.002$ for NoScooter group; $r=-0.32$, $p=0.011$ for Scooter group. At the same time, carelessness was positively correlated with engaged concentration, $r=0.58$, $p<0.001$ for NoScooter group; $r=0.54$, $p<0.001$ for Scooter group. This indicated that more engaged students were careless more often, while less engaged students were careless less often. Carelessness and frustration were not significantly correlated in either condition, $r=-0.18$, $p=0.17$ for NoScooter group; $r=0.13$, $p=0.32$ for Scooter group. In the following analyses, we will explore these three significant relationships in greater detail.

We examined the changes in carelessness and affect over time to see if there were significant differences in the relationship and occurrence of carelessness and affect as the student used the tutor. We did this by separating the observed affect and estimates of carelessness during the student's tutor usage into two halves. For the observed affect, the first 12 affect observations for each student (across both raters) were grouped as the first half, and the remaining 12 observations per student were grouped as the second half. To split the estimates of carelessness into halves, we took and split the overall time of each student's usage of the tutor (splitting by time rather than number of actions), and grouped the first half and second half of that student's actions based on the total elapsed time at each action. We computed the average carelessness for incorrect actions during each half of each student's tutor usage. Figure 5 shows the student affect percentages and average estimates of carelessness in each half of the students' tutor usage in both NoScooter and Scooter groups.

Across time, the changes in the incidence of different affective states were consistent across conditions. In both conditions, students exhibited less confusion in the second half of their usage than the first half, $F(1,124)=61.76$, $p<0.001$. They also exhibited more engaged concentration over time, $F(1,124)=40.26$, $p<0.001$, and exhibited more boredom over time, $F(1,124)=11.89$, $p=0.001$. The changes in the proportion of these affective states over time were not significantly different between the NoScooter and Scooter conditions – for boredom $F(1,124)=0.01$, $p=0.93$, for confusion $F(1,124)=0.79$, $p=0.38$, and for engaged concentration $F(1,124)=1.82$, $p=0.18$.

For carelessness, the degree was significantly different during the second half of tutor usage across conditions, $F(1,124)=22.85$, $p<0.001$). However, carelessness

Table 3 Correlations of carelessness and affective state for entire tutor usage

	NoScooter group	Scooter group
Careless – Boredom	**−0.36 ($p=0.004$)**	**−0.35 ($p=0.004$)**
Careless – Confusion	**−0.38 ($p=0.002$)**	**−0.32 ($p=0.011$)**
Careless – Engaged Concentration	**+0.58 ($p<0.001$)**	**+0.54 ($p<0.001$)**
Careless – Frustration	−0.18 ($p=0.17$)	+0.13 ($p=0.32$)

Statistically significant correlations in bold

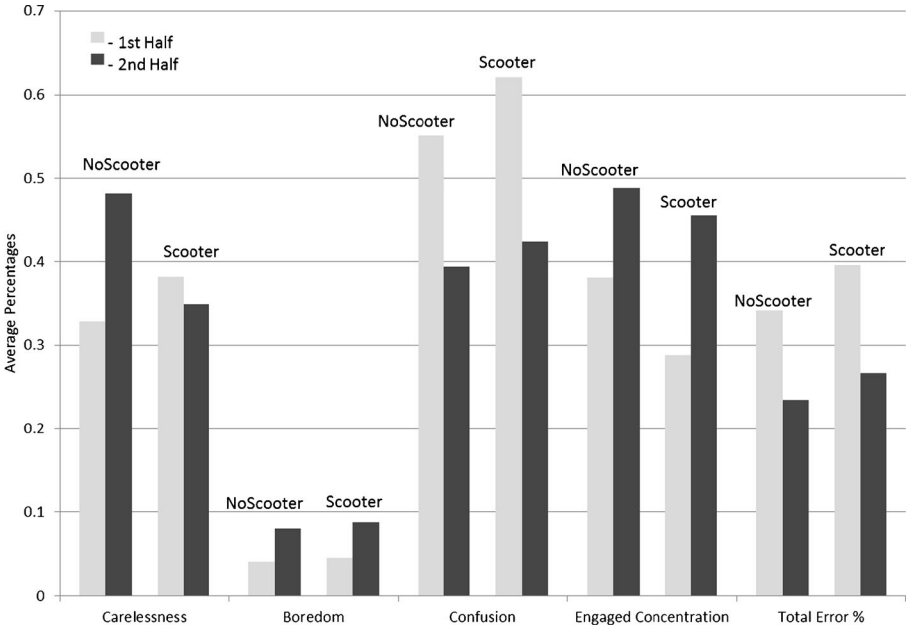

Fig. 5 Carelessness, affect, error, by time of tutor usage

showed a significant trend towards increase over time in the NoScooter condition, $t(61)=-8.23$, two-tailed $p<0.001$, while a marginally significant decrease over time in the Scooter condition, $t(63)=1.88$, two-tailed $p=0.06$. The difference in trend between conditions (e.g. the interaction effect) was statistically significant, $F(1,124)=53.81$, $p<0.001$. Overall, student errors (whether careless or not) significantly decreased over time in both conditions, $F(1,124)=74.53$, $p<0.001$. This effect was seen in both the NoScooter condition ($t(61)=5.70$, two-tailed $p<0.001$) and the Scooter condition ($t(63)=6.50$, two-tailed $p<0.001$). The changes in the number of errors over time were not significantly different between the NoScooter and Scooter conditions, $F(1,124)=0.63$, $p=0.43$. This, together with the trend in carelessness, may suggest that Scooter was effective overall in reducing student errors, including both careless and non-careless errors.

Table 4 shows the correlations between carelessness and the three most common affective states over time. During the first half of the tutor usage period, the correlations between carelessness and boredom trended negative in both groups, but were not significant, $r=-0.19$, $p=0.15$ for NoScooter group; $r=-0.09$, $p=0.47$ for Scooter group. During the second half, the correlations became significantly negative, $r=-0.32$, $p=0.01$ for NoScooter group; $r=-0.34$, $p=0.01$ for Scooter group. Steiger's Z-test (1980), a standard test for comparing multiple correlations with no overlapping variables within a single population, was used to determine whether the correlation between carelessness and boredom was significantly different between the two time periods. The difference though in correlation was not significant for both the NoScooter group, $Z=0.94$, $p=0.34$, and the Scooter group, $Z=1.72$, $p=0.08$.

During the first half of the tutor usage period, the correlations between carelessness and confusion trended negative in both groups, but were not significant, $r=-0.07$, $p=0.57$ for NoScooter group; $r=-0.18$, $p=0.16$ for Scooter group. During the second

Table 4 Correlations between carelessness and affect by time within tutor usage

	NoScooter group	Scooter group
Careless – Boredom (1st Half)	−0.19 (p=0.15)	−0.09 (p=0.47)
Careless – Confusion (1st Half)	−0.07 (p=0.57)	−0.18 (p=0.16)
Careless – Engaged Concentration (1st Half)	+0.17 (p=0.20)	**+0.31 (p=0.014)**
Careless – Boredom (2nd Half)	**−0.32 (p=0.012)**	**−0.34 (p=0.006)**
Careless – Confusion (2nd Half)	**−0.37 (p=0.004)**	**−0.33 (p=0.009)**
Careless – Engaged Concentration (2nd Half)	**+0.55 (p<0.001)**	**+0.50 (p<0.001)**

Significant corrections in bold

half, the correlations became significantly negative, r=−0.37, p=0.004 for NoScooter group; r=−0.33, p=0.009 for Scooter group. However, the difference in correlation between the time periods was not statistically significant for either group, Z=1.83, p=0.07 for NoScooter group; Z=0.97, p=0.33 for Scooter group.

During the first half of the tutor usage period, the correlations between carelessness and engaged concentration trended positive in both groups, but were not significant for NoScooter group, r=0.17, p=0.20. For the Scooter group, however, tutor usage for the first half was significant, r=0.31, p=0.014. During the second half, the correlations became significantly positive, r=0.55, p<0.001 for NoScooter group; and remained significant, r=0.50, p<0.001, for Scooter group. The correlation between carelessness and engaged concentration was significantly different between the first half of tutor usage and the second half of tutor usage, for NoScooter group, Z=−2.73, p=0.006; but not significantly different for Scooter group, Z=−1.48, p=0.14.

We can see here that the relationship between carelessness and affective states were mostly significant during the second half of tutor usage. Looking at the Steiger's Z values, however, shows that carelessness' positive or negative relationship with the affective states have been fairly consistent from start to end. Further discussion about these relationships over time will be given below.

Discussion and Conclusion

Carelessness has been identified as a problematic behavior in classrooms since the 1950s (e.g. Eaton et al. 1956). Within the context of educational software, the disengaged behavior of carelessness has not been heavily studied compared to other student behaviors. In this paper, we leverage educational data mining to model carelessness, towards understanding the factors that lead a student to be careless in a given situation. More specifically, we study the relationship between student carelessness and affective states within a Cognitive Tutor for Scatterplots, using automated detection of carelessness and field observations of student affect. Student carelessness, operationalized as the probability that an error was due to slipping, was estimated using the Contextual Slip model by Baker et al. (2008b), and implemented for this specific tutor. The detector infers carelessness from the features of the individual student action.

The detector's assessment of each student's carelessness was then studied in conjunction with that student's proportion of each affective state. Several relationships

were found. First, the more confused a student is, the less likely errors are to be careless. The negative relationship between confusion and carelessness is reasonable. In particular, not knowing a skill is likely to be associated with confusion (cf. D'Mello et al. 2009; Baker et al. 2010; Rodrigo et al. 2010; Lee et al. 2011), and not knowing a skill will increase the number of errors that cannot be attributed to carelessness (as our model requires that a skill to be known for an error to be seen as careless). As such, this finding accords to what might be expected. Second, the more bored a student is, the less likely errors are to be careless. The significant negative correlation between carelessness and boredom is somewhat more surprising – for example, Cheyne et al. (2006) hypothesize that boredom proneness leads to carelessness. One possible explanation can be attributed to previous findings that boredom is correlated with poorer learning (Craig et al. 2004; Pekrun et al. 2010), and the hypothesis that boredom leads to shallow information processing and reduce the use of task-related cognitive strategies (Pekrun et al. 2010). As such, boredom may have led to lower knowledge, which in turn would have led to more errors that cannot be attributed to carelessness. Third, the more a student displays the affective state of engaged concentration, the more likely he/she is to display carelessness. One possible hypothesis for this finding comes from work by Clements (1982), which found a positive relationship between carelessness and confidence, such that students who understood the material and believed they knew how to answer (or even answered relatively quickly), made a higher proportion of careless errors. Hence, a student with average knowledge who is engaged may have tendencies to overestimate their own ability (cf. Prohaska and Maraj 1995; Linnenbrink and Pintrich 2003), leading them to become careless.

Examination of slip and affective state behavior over time and their respective correlations supported these hypotheses. The correlation between confusion and carelessness became significantly negative as students used the tutor more. However, confusion (as well as total errors – whether careless or not) decreased over time. Hence, the students who were struggling most and remained confused even after using the tutor for a substantial amount of time were less likely to make errors that were inferred as careless. Similarly, the correlation between carelessness and engaged concentration became positive and significant predominantly during the second half of tutor usage (in NoScooter group, but in both halves for the Scooter group). One possible interpretation for this is that a student who is engaged most of the time may succeed, become overconfident, and then commit careless errors. It is interesting to note that the connections between affect and carelessness were generally substantially stronger in the second half of tutor usage than the first, across affective states. It is possible that as the students used the tutor more, careless errors became more and more driven by the presence of affect (or decrease of it, as in confusion), even as the student learned more over time.

Another possible interpretation, however, is that the careless errors may instead be due to the relatively low cost to the learner from making a mistake. In other words, once a student generally knows the answers, he or she may choose to work quickly, knowing that some mistakes may occur, but that the consequences will be limited. The learner may even be testing him or herself to see how fast he or she can complete the problems, an engaging and challenging activity (which would explain the correlation between carelessness and engaged concentration, as well as the relationship to mastery goals), leading to some seemingly careless errors. A strongly engaged user need not be reflecting deeply about their work.

Overall, these patterns seem very different from those seen for other forms of disengagement. For example, previous research has found positive associations between

boredom and gaming the system (Baker et al. 2010b), off-task behavior (Baker et al. 2011), and using the software in a fashion unrelated to the educational task (Sabourin et al. 2011). This suggests that the motivations underlying a student's carelessness may be quite different from the motivations underlying other disengaged behaviors.

It is worth noting that there have also been surprising relationships found between carelessness and goal orientation (Hershkovitz et al. 2011). In this work, carelessness was found to be positively correlated with academic efficacy and negatively correlated with disruptive behavior and self-presentation of low achievement (Hershkovitz et al. 2011). In addition, carelessness was found to be higher among students with mastery or performance goals than among students manifesting neither type of goal orientation.

Thus, it is worth noting that the factors carelessness is associated with are typically thought to be associated with more successful students. For instance, engaged concentration has been repeatedly shown to be associated with successful learning (Craig et al. 2004), and boredom has been repeatedly shown to be associated with poorer learning (Craig et al. 2004; Pekrun et al. 2010). In addition, mastery and performance goals have been shown to be associated with successful learning (Harackiewicz et al. 2002), as is self-efficacy (Zimmerman et al. 1992). As such, it may be that carelessness will turn out to be a form of disengaged behavior characteristic of generally more successful students, a pattern also seen in Clements (1982) and Baker et al. (2010). As such, the previous notions of careless students as lazy (Hurlock and McDonald 1934), absent-minded (Eysenck and Keane 1990), and unorganized (Epstein 1979) may not be the most appropriate description for careless students.

Instead, we may need to consider carelessness as a risk to engagement for students who are generally successful. These results illustrate the key role of affect in student carelessness, and suggest that adaptive responses to carelessness should take probable student affect into account. Based on the evidence from this study, carelessness is a risk for students who are typically highly engaged. As such, it may be appropriate to tailor messages given to careless students to the motivational needs of highly engaged students (Graesser et al. 2008), potentially leading to very different intervention than is typically given to disengaged students. While these highly engaged students are typically successful, they are still at risk of becoming disengaged (Csikszentmihalyi and Schneider 2000), or in having consequences due to their careless errors, especially in high-stakes situations such as standardized exams. In particular, though carelessness is generally characteristic of successful and engaged students, it is nonetheless associated with failing to attend college, once student knowledge is controlled for (San Pedro et al. 2013). As such, it remains important to understand carelessness better and address it effectively, in order to move towards computer-aided instruction and computer-aided learning which effectively addresses the disengagement of all students.

Acknowledgments This research was supported by the Pittsburgh Science of Learning Center (National Science Foundation) via grant "Toward a Decade of PSLC Research", award number SBE-0836012, and the Philippines Department of Science and Technology Philippine Council for Advanced Science and Technology Research and Development under the project "Development of Affect-Sensitive Interfaces". We thank Mrs. Carmela Oracion, Jenilyn Agapito, Ivan Jacob Pesigan, Ma. Concepcion Repalam, Salvador Reyes, Ramon Rodriguez, 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.

References

Adolphs, R., & Damasio, A. R. (2001). The interaction of affect and cognition: A neurobiological perspective. In J. P. Forgas (Ed.), *Handbook of affect and social cognition* (pp. 27–49). Mahwah: Lawrence Erlbaum Associates.

Aleven, V., McLaren, B. M., Roll, I., & Koedinger, K. R. (2004). Toward tutoring help seeking: Applying cognitive modeling to meta-cognitive skills. In J. C. Lester, R. M. Vicari, & F. Paraguacu (Eds.), *Proceedings of the 7th International Conference on Intelligent Tutoring Systems* (pp. 227–239). Berlin: Springer.

Aleven, V., McLaren, B. M., Roll, I., & Koedinger, K. R. (2006). Toward meta-cognitive tutoring: a model of help seeking with a cognitive tutor. *International Journal of Artificial Intelligence in Education, 16*, 101–130.

Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., & Christopherson, R. (2009). Emotion Sensors Go To School. *Proceedings of the International Conference on Artificial Intelligence in Education*, 17–24.

Baker, R. S. J. d., & Gowda, S. M. (2010). An analysis of the differences in the frequency of students' disengagement in urban, rural, and suburban high schools. In R. S. J. d. Baker, A. Merceron & P. I. Pavlik (Eds.), *Proceedings of the 3rd International Conference on Educational Data Mining* (pp. 11–20).

Baker, R. S. J. d., & Yacef, K. (2009). The state of educational data mining in 2009: a review and future visions. *Journal of Educational Data Mining, 1*, 3–17.

Baker, R. S. J. d., Corbett, A. T., Koedinger, K. R., & Wagner A. Z. (2004). Off-task behavior in the cognitive tutor classroom: when students "game the system". In *Proceedings of ACM CHI 2004: Computer-Human Interaction* (pp. 383–390). ACM New York, NY.

Baker, R. S. J. d., Corbett, A. T., Koedinger, K. R., Evenson, S. E., Roll, I., Wagner, A. Z., Naim, M., Raspat, J., Baker, D. J., & Beck, J. (2006). Adapting to when students game an intelligent tutoring system. In M. Ikeda, K. D. Ashley, & T. W. Chan (Eds.), *Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (pp. 392–401). Berlin: Springer Verlag.

Baker, R. S. J. d., Corbett, A. T., & Aleven, V. (2008a). Improving contextual models of guessing and slipping with a truncated training set. In R. S. J. d. Baker & J. Beck (Eds.), *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 67–76).

Baker, R. S. J. d., Corbett, A. T., & Aleven, V. (2008b). More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In E. Aimeur & B. Woolf (Eds.), *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (pp. 406–415). Berlin: Springer Verlag.

Baker, R. S. J. d., Walonoski, J. A., Heffernan, N. T., Roll, I., Corbett, A. T., & Koedinger, K. R. (2008c). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research, 19*, 185–224.

Baker, R. S. J. d., Corbett, A. T., Gowda, S. M., Wagner, A. Z., MacLaren, B. M., Kauffman, L. R., Mitchell, A. P., & Giguere, S. (2010a). Contextual slip and prediction of student performance after use of an intelligent tutor. In P. De Bra, A. Kobsa, & D. Chin (Eds.), *Proceedings of the 18th Annual Conference on User Modeling, Adaptation, and Personalization* (pp. 52–63). Berlin: Springer Verlag.

Baker, R. S. J. d., D'Mello, S. K., Rodrigo, M. M. T., & Graesser, A. C. (2010b). 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, 68*, 223–241.

Baker, R. S. J. d., Moore, G., Wagner, A., Kalka, J., Karabinos, M., Ashe, C. & Yaron, D. (2011). The dynamics between student affect and behavior occuring outside of educational software. In *Proceedings of the 4th bi-annual International Conference on Affective Computing and Intelligent Interaction*, pp. 14–24.

Balfanz, R., Herzog, L., & Mac, I. D. J. (2007). Preventing student disengagement and keeping students on the graduation path in urban middle-grade schools: early identification and effective interventions. *Educational Psychologist, 42*, 223–235.

Bartel, C. A., & Saavedra, R. (2000). The collective construction of work group moods. *Administrative Science Quarterly, 45*(2), 197–231.

Becker, H. J. (2000). Pedagogical motivations for student computer use that lead to student engagement. *Educational Technology, 40*, 5–17.

Cheyne, J. A., Carriere, J. S., & Smilek, D. (2006). Absent-mindedness: lapses of conscious awareness and everyday cognitive failures. *Consciousness and Cognition, 15*(3), 578–592.

Clements, K. (1982). Careless errors made by sixth-grade children on written mathematical tasks. *Journal for Research in Mathematics Education, 13*, 136–144.

Cocea, M., Hershkovitz, A., & Baker, R. S. J. d. (2009). The impact of off-task and gaming behaviors on learning: immediate or aggregate? In *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 507–514). Amsterdam: IOS Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction, 4*, 253–278.

Craig, S. D., Graesser, A. C., Sullins, J., & Gholson, B. (2004). Affect and learning: an exploratory look into the role of affect in learning with autotutor. *Journal of Educational Media, 29*, 241–250.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York: Harper and Row.

Csikszentmihalyi, M., & Schneider, B. (2000). *Becoming adult*. New York: Basic Books.

D'Mello, S. K., & Graesser, A. C. (2012). Dynamics of affective states during complex learning. *Learning and Instruction, 22*, 145–157.

D'Mello, S. K., Craig, S. D., Sullins, J., & Graesser, A. C. (2006). Predicting affective states expressed through an emote-aloud procedure from AutoTutor's mixed-initiative dialogue. *International Journal of Artificial Intelligence in Education, 16*(1), 3–28.

D'Mello, S., Graesser, A., & Picard, R. W. (2007). Toward an affect-sensitive AutoTutor. *IEEE Transactions on Intelligent Systems, 22*(4), 53–61.

D'Mello, S. K., Person, N., & Lehman, B. (2009). Antecedent-consequent relationships and cyclical patterns between affective states and problem solving outcomes. In V. Dimitrova, R. Mizoguchi, B. du Buolay, & A. C. Graesser (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 57–64). Amsterdam: Ios Press.

D'Mello, S. K., Lehman, B., & Person, N. (2010). Monitoring affect states during effortful problem solving activities. *International Journal of Artificial Intelligence in Education, 20*(4), 361–389.

Eaton, M. T., D'Amico, L. A., & Phillips, B. N. (1956). Problem behavior in school. *Journal of Educational Psychology, 47*, 350–357.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician, 37*, 36–48.

Ekman, P., & Friesen, W. V. (1978). *The facial action coding system*. Palo Alto: Consulting Psychologists Press.

Epstein, S. (1979). The stability of behavior: I. on predicting most of the people much of the time. *Journal of Personality and Social Psychology, 37*, 1097–1126.

Eysenck, M. W., & Keane, M. T. (1990). *Cognitive psychology: A Student's handbook*. London: Lawrence Erlbaum Associates.

Forster, M. R. (2000). Key concepts in model selection: performance and generalizability. *Journal of Mathematical Psychology, 44*, 205–231.

Fredrickson, B. L., & Branigan, C. (2005). Positive emotions broaden the scope of attention and thought-action repertoires. *Cognition and Emotion, 19*, 313–332.

Graesser, A. C., McDaniel, B., Chipman, P., Witherspoon, A., D'Mello, S., & Gholson, B. (2006). Detection of emotions during learning with AutoTutor. *Proceedings of the 28th Annual Meeting of the Cognitive Science Society* (pp. 285–290).

Graesser, A. C., Rus, V., D'Mello, S., & Jackson, G. T. (2008). AutoTutor: learning through natural language dialogue that adapts to the cognitive and affective states of the learner. In D. H. Robinson & G. Schraw (Eds.), *Current perspectives on cognition, learning and instruction: Recent innovations in educational technology that facilitate student learning* (pp. 95–125). Charlotte: Information Age Publishing.

Harackiewicz, J. M., Barron, K. E., Tauer, J. M., & Elliot, A. J. (2002). Predicting success in college: a longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. *Journal of Educational Psychology, 94*, 562–575.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Informatics and Computer Science, 44*, 1–12.

Hay, J. F., & Jacoby, L. L. (1996). Separating habit and recollection: memory slips, process dissociations and probability matching. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22*, 1323–1335.

Heffernan, N., Koedinger, K., & Razzaq, L. (2008). Expanding the model-tracing architecture: a 3rd generation intelligent tutor for algebra symbolization. *The International Journal of Artificial Intelligence in Education, 18*(2), 153–178.

Hershkovitz, A., Wixon, M., Baker, R. S. J. d., Gobert, J., & Sao Pedro, M. (2011). Carelessness and goal orientation in a science microworld. In J. Kay, S. Bull, & G. Biswas (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 462–465). Berlin: Springer.

Hurlock, E. B., & McDonald, L. C. (1934). Undesirable behavior traits in junior high school students. *Child Development, 5*, 278–290.

Karweit, N., & Slavin, R. (1981). Measurement and modeling choices in studies of time and learning. *American Educational Research Journal, 18*, 157–171.

Koedinger, K. R., & Corbett, A. T. (2006). Cognitive tutors: Technology bringing learning sciences to the classroom. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning science* (pp. 61–77). New York: Cambridge University Press.

Koedinger, K. R., Baker, R. S. J. d., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the edm community: The pslc datashop. In C. Romero, C. Ventura, M. Pechenizkiy, & R. S. J. d. Baker (Eds.), *Handbook of educational data mining)* (pp. 43–56). Boca Raton: CRC Press.

Kort, B., Reilly, R., & Picard, R. (2001). An affective model of interplay between emotions and learning: Reengineering educational pedagogy—building a learning companion. In T. Okamoto, R. Hartley, Kinshuk, & J. P. Klus (Eds.), *IEEE Proceedings International Conference on Advanced Learning Technology: Issues, Achievements and Challenges* (pp. 43–48). Madison: IEEE Computer Society.

Lau, S., & Darmanegara, L. A. (2007). The role of self-efficacy, task value, and achievement goals in predicting learning strategies, task disengagement, peer relationship, and achievement outcome. *Contemporary Educational Psychology, 3*, 1–26.

Lee, D. M., Rodrigo, M. M. T., Baker, R. S. J. d., Sugay, J., & Coronel, A. (2011). Exploring the relationship between novice programmer confusion and achievement. In S. K. D'Mello, A. C. Graesser, B. Schuller, & J. Martin (Eds.), *Proceedings of the 4th bi-annual International Conference on Affective Computing and Intelligent Interaction* (pp. 175–184). Berlin Heidelberg: Springer.

Linnenbrink, E. A., & Pintrich, P. R. (2003). The role of self-efficacy beliefs in student engagement and learning in the classroom. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 19*, 119–137.

Mandler, G. (1984). *Mind and body: Psychology of emotion and stress*. New York: W.W. Norton & Company.

Maydeu-Olivares, A., & D'Zurilla, T. J. (1996). A factor-analytic study of the social problem-solving inventory: an integration of theory and data. *Cognitive Therapy and Research, 20*, 115–133.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). YALE: Rapid prototyping for complex data mining tasks. In L. Ungar, M. Craven, D. Gunopulos, & T. Eliassi-Rad (Eds.), *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 935–940). New York: ACM.

Mitrovic, A. (2003). An intelligent SQL tutor on the web. *International Journal of Artificial Intelligence in Education, 13*(2), 173–197.

Mitrovic, A., Martin, B., Suraweera, P., Zakharov, K., Milik, N., Holland, J., & McGuigan, N. (2009). ASPIRE: an authoring system and deployment environment for constraint-based tutors. *International Journal of Artificial Intelligence in Education, 19*(2), 155–188.

Myung, I. J., Pitt, M. A., & Kim, W. (2005). Model evaluation, testing and selection. In K. Lambert & R. Goldstone (Eds.), *The handbook of cognition* (pp. 422–436). Thousand Oaks: Sage.

National Research Council and Institute of Medicine. (2004). *Engaging schools: Fostering high school Students' motivation to learn*. Washington, DC: The National Academies Press.

Newman, M. (1977). An analysis of sixth-grade pupils' errors on written mathematical tasks. *Victorian Institute for Educational Research Bulletin, 39*, 31–43.

Norman, D. (1981). Categorization of action slips. *Psychological Review, 88*, 1–15.

Nottelmann, E. D., & Hill, K. T. (1977). Test anxiety and off-task behavior in evaluative situations. *Child Development, 48*, 225–231.

Ocumpaugh, J., Baker, R. S. J. d., & Rodrigo, M. M. T. (2012) Baker-Rodrigo *Observation Method Protocol (BROMP) 1.0. Training Manual version 1.0. Technical Report.* New York, NY: EdLab. Manila, Philippines: Ateneo Laboratory for the Learning Sciences.

Pekrun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., & Perry, R. P. (2010). Boredom in achievement settings: exploring control–value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology, 102*, 531–549.

Planalp, S., DeFrancisco, V. L., & Rutherford, D. (1996). Varieties of cues to emotion in naturally occurring settings. *Cognition and Emotion, 10*(2), 137–153.

Prohaska, V., & Maraj, F. (1995). Low and medium ability students confidently overestimate all their grades. Presented at the *Seventh Annual Convention of the American Psychological Society.* New York.

Razzaq, L. M., & Heffernan, N. T. (2009, July). To Tutor or Not to Tutor: That is the Question. In *AIED* (pp. 457–464).

Razzaq, L. M., Feng, M., Nuzzo-Jones, G., Heffernan, N. T., Koedinger, K. R., Junker, B., & Rasmussen, K. P. (2005, May). Blending Assessment and Instructional Assisting. In *AIED* (pp. 555–562).

Rodrigo, M. M. T., Baker, R. S. J. d., & Nabos, J. (2010). The relationships between sequences of affective states and learner achievement. In S. L. Wong, S. C. Kong, & F. Y. Yu (Eds.), *Proceedings of the 18th International Conference on Computers in Education* (pp. 56–60). Malaysia: Universiti Putra Malaysia.

Rodrigo, M. M. T., Baker, R. S. J. d., Agapito, J., Nabos, J., Repalam, M. C., Reyes, S. S., et al. (2012). The Effects of an Interactive Software Agent on Student Affective Dynamics while Using an Intelligent Tutoring System. *IEEE Trans. Affective Computing*, 3, 2.

Rosenthal, R., & Rosnow, R. L. (2008). *Essentials of behavioral research: Methods and data analysis*. New York: McGraw-Hill Humanities.

Rowe, J., McQuiggan, S., Robison, J., & Lester, J. (2009). Off-task behavior in narrative-centered learning environments. In V. Dimitrova, R. Mizoguchi, B. du Buolay, & A. C. Graesser (Eds.), *Proceedings of the 14th International Conference on Artificial Intelligence in Education* (pp. 99–106). Amsterdam: Ios Press.

Rumberger, R. W., & Larson, K. A. (1998). Student mobility and the increased risk of high school dropout. *American Journal of Education, 107*, 1–35.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review, 110*, 145–172.

Ryan, A. M., & Patrick, H. (2001). The classroom social environment and changes in adolescents' motivation and engagement during middle school. *American Educational Research Journal, 38*, 437–460.

Sabourin, J., Rowe, J., Mott, B., & Lester, J. (2011). When off-task in on-task: The affective role of off-task behavior in narrative-centered learning environments. In J. Kay, S. Bull, & G. Biswas (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 534–536). Berlin: Springer.

San Pedro, M. O. C., Baker, R. S. J. d., Rodrigo, M. M. (2011a) The Relationship between Carelessness and Affect in a Cognitive Tutor. In *Proceedings of the 4th bi-annual International Conference on Affective Computing and Intelligent Interaction* (pp. 306–315).

San Pedro, M. O. C., Baker, R., Rodrigo, M. M. (2011b) Detecting Carelessness through Contextual Estimation of Slip Probabilities among Students Using an Intelligent Tutor for Mathematics. In *Proceedings of 15th International Conference on Artificial Intelligence in Education* (pp. 304–311).

San Pedro, M. O. Z., Baker, R. S. J. d., Bowers, A. J., & Heffernan, N. T. (2013) Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School. *Proceedings of the 6th International Conference on Educational Data Mining* (pp. 177–184).

Schofield, J. W. (1995). *Computers and classroom culture*. Cambridge: Cambridge University Press.

Schutzwohl, A., & Borgstedt, K. (2005). The processing of affectively valenced stimuli: the role of surprise. *Cognition & Emotion, 19*, 583–600.

Soriano, J. C. A., Rodrigo, M. M. T., Baker, R. S. J. d., Ogan, A., Walker, E., Castro, M. J., et al. (2012) A Cross-Cultural Comparison of Effective Help Seeking Behavior among Students Using an ITS for Math. Poster paper. *Proceedings of the International Conference on Intelligent Tutoring Systems* (pp. 636–637).

Sosa, G., Berger, D. E., Saw, A. T., & Mary, J. C. (2011). Effectiveness of computer-based instruction in statistics: a meta-analysis. *Review of Educational Research, 81*, 97–128.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin, 87*, 245–251.

Suraweera, P., & Mitrovic, A. (2004). An intelligent tutoring system for entity relationship modelling. *International Journal of Artificial Intelligence in Education, 14*(3), 375–417.

Wixon, M., Baker, R. S. J. d., Gobert, J., Ocumpaugh, J., & Bachmann, M. (2012) WTF? Detecting Students who are Conducting Inquiry Without Thinking Fastidiously. *Proceedings of the 20th International Conference on User Modeling, Adaptation and Personalization (UMAP 2012)* (pp. 286–298).

Woolf, B., Burleson, W., Arroyo, I., Dragon, T., Cooper, D., & Picard, R. (2009). Affect-aware tutors: recognising and responding to student affect. *International Journal of Learning Technology, 4*(3), 129–164.

Zimmerman, B. J., Bandura, A., & Martinez-Pons, M. (1992). Self-motivation for academic attainment: the role of self-efficacy beliefs and personal goal setting. *American Educational Research Journal, 29*, 663–676.