# Detecting Carelessness through Contextual Estimation of Slip Probabilities among Students Using an Intelligent Tutor for Mathematics

Maria Ofelia Clarissa Z. SAN PEDRO[1], Ryan S. d. J. BAKER[2], Ma. Mercedes T. RODRIGO[1]
[1] *Ateneo de Manila University, Loyola Heights, Quezon City, Philippines*
[2] *Worcester Polytechnic Institute, Worcester, MA*
*sweetsp@gmail.com, rsbaker@wpi.edu, mrodrigo@ateneo.edu*

**Abstract.** A student is said to have committed a careless error when a student's answer is wrong despite the fact that he or she knows the answer (Clements, 1982). In this paper, educational data mining techniques are used to analyze log files produced by a cognitive tutor for Scatterplots to derive a model and detector for carelessness. Bayesian Knowledge Tracing and its variant, the Contextual-Slip-and-Guess Estimation, are used to model and predict carelessness behavior in the Scatterplot Tutor. The study examines as well the robustness of this detector to a major difference in the tutor's interface, namely the presence or absence of an embodied conversational agent, as well as robustness to data from a different school setting (USA versus Philippines).

**Keywords:** Carelessness, Slip, Contextual-Slip-and-Guess, Bayesian Knowledge Tracing, Cognitive Tutors, Scatterplot.

## 1 Introduction

Recently, there has been increasing attention to studying disengaged behaviors within intelligent tutoring systems [2, 6]. One student behavior that has been less thoroughly explored is carelessness [8, 9, 10] – a label ascribed to the unconscientious performance of actions that were not originally intended by the individual, usually leading to errors [13, 15]. This can happen when an individual is in a hurry or overconfident in carrying out a task, when doing routine activities, or when doing tasks perceived to be of minor importance [12]. Carelessness is not an uncommon behavior in students [8], even among high-performing students [9]. Modeling this student behavior may lead not only to a fuller understanding of a student's true learning capabilities, but also to improved teaching strategies and educational materials.

Recent studies have shown educational software to be useful in measuring student affect, knowledge, and disengaged behavior within a classroom setting. One type of educational software, an Intelligent Tutoring System (ITS), provides students with guided learning support as they engage in problem-solving [16]. Researchers have used ITSs in modeling student learning, approximating the knowledge state of each

student at a given time [11]. In recent years, further studies using ITSs have branched out towards modeling and detecting student affective states [1, 17] and behaviors associated with affect and poorer learning, including gaming the system [6] and off-task behavior [2]. Of importance to the analyses in this paper, Baker, et al. [4] have recently developed a slip detector [4] which can be used to detect carelessness as student behavior within ITSs. This operationalization of carelessness accords to the definition of carelessness in Clements, that errors committed by students deemed competent in problem-solving indicate carelessness behavior [9]. However, although the model has been applied within multiple tutors, it is not yet clear how widely the model generalizes. For this model to be broadly useful, it must be able to generalize to new tutor designs and student populations.

Within this paper, we establish the generalizability of models of students' carelessness, using two versions of a Cognitive Tutor for Scatterplot generation and interpretation, differing in the presence or absence of an Embodied Conversational Agent (ECA) [6]. We analyze interaction logs from Philippine high school students under these two conditions, producing two slip detectors based on previous work at modeling this construct [3, 4]. We then test the detectors on the other version of the learning environment's dataset to see how well the detectors generalize to data sets with significant differences in design. We also test the detectors on interaction logs from US middle school students using the same tutors to see how well these models generalize to data with a different school setting. In the long term, the work hopes to contribute to a generalizable model of carelessness.

## 2 Carelessness Detection in Cognitive Tutors

Cognitive Tutors employ a strategy known as Knowledge Tracing to estimate a student's latent knowledge based on his/her observable performance. This process is based on Corbett and Anderson's Bayesian Knowledge Tracing (BKT) model [11].

The BKT framework, in its original articulation, enables the Cognitive Tutor 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 regardless whether the response is correct or not. It uses four parameters – two learning parameters $L_O$ (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) – for each skill (estimated from data information in each skill). These parameters are invariant across the entire context of using the tutor. 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 [11]:

$$P(L_n \mid Action_n) = P(L_{n-1} \mid Action_n) + ((1 - P(L_{n-1} \mid Action_n)) * P(T)) \quad . \qquad (1)$$

Studies by Baker et al. proposed a variant of the BKT model which contextually estimates the Guess and Slip parameters, with this Contextual Slip being an indicator of carelessness [3, 4]. The Contextual Guess-and-Slip (CGS) model examines the properties of each student response as it occurs, in order to assess the probability that the response to an action is a guess or slip. In this model, the estimates of the slip and guess probabilities are now dynamic and depends on the contextual information of the action, such as speed and history of help-seeking from the tutor. It has been shown that this model can indicate aspects of student learning that are not captured by traditional BKT, which may significantly improve prediction of post-test performance [5]. Based on prior theory on carelessness (as discussed above), we use the slip model as an operationalization of carelessness [cf. 8] (though slips may also occur for other reasons, such as shallow knowledge [e.g. 5]).

## 3 Methods

Data were gathered from 126 students from a large public high school in Quezon City, Philippines (PH). For 80 minutes, students used a Cognitive Tutor unit on scatterplot generation and interpretation [6]. Students had not explicitly covered these topics in class prior to the study. Prior to using the software, students viewed conceptual instruction. Each student in each class took a nearly isomorphic pre-test and post-test, counterbalanced across conditions.

Within the Scatterplot Tutor, 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 is monitored and displayed through skill bars that depict his/her mastery of skills.

Sixty four of the participants (Scooter group) were randomly assigned to use a version of the tutor with an embodied conversational agent, "Scooter the Tutor". Scooter was designed to both reduce the incentive to prevent gaming the system and to help students learn the material that they were avoiding by gaming, while affecting non-gaming students as minimally as possible. Gaming the system is defined in [6] as behavior aimed at obtaining correct answers and advancing within the tutoring curriculum by systematically taking advantage of regularities in the software's feedback and help. Scooter displays happiness and gives positive message when students do not game (regardless of the correctness of their answers), but shows dissatisfaction when students game, and provides supplementary exercises to help them learn material bypassed by gaming. The remaining 62 participants (NoScooter group) used a version of the Scatterplot Tutor without the conversational agent. As such, skills associated with the tutor version with Scooter have additional Scooter-related skills not present in the tutor without Scooter. The number of students assigned to the conditions in this study was unbalanced because of data gathering schedule disruptions caused by inclement weather.

Log files generated by the Cognitive tutor recorded the students' actions in real-time. A set of 26 transaction features identical to the set used in [4] was extracted and derived from the logs for each problem step. These features were used since they have been shown to be effective in creating detectors of other constructs [e.g. 6]. Baseline BKT parameters were fit with brute-force search [cf. 5]. From this baseline model, estimates of whether the student knew the skill at each step were derived and used to label actions (whether correct or incorrect response) with the probability that the actions involved guessing or slipping, based on the student performance on successive opportunities to apply the rule [4]. As in [3, 4], Bayesian equations were utilized in computing training labels for the Slip (and Guess) probabilities for each student action (A) at time N, using future information (two actions afterwards – N+1, N+2), in order to infer as accurately as possible the true probability that a student's action at time N was due to knowing the skill, or due to a slip or guess [4]. Using Eq. 2, 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 ($A_{N+1,N+2}$).

$$P(A_N \text{ is a Slip} \mid A_N \text{ is incorrect}) = P(L_n \mid A_{N+1,N+2}) . \quad (2)$$

Models for Contextual Slip (and Guess) were then produced through Linear Regression using truncated training data [3], to create models that could predict contextual guess and slip without using data from the future. These new models were then substituted for the Guess and Slip parameters per problem step, labeling each action with variant estimates as to how likely the response is a guess or a slip. With dynamic values of Guess/Slip, the learning parameters Lo and T were re-fit per skill.

## 4 Results and Discussion

Using student-level cross-validation (6-fold) Linear Regression Modeling in RapidMiner, a Carelessness model approximating the Contextual Slip Model was created with the 26 attributes extracted, plus the label of the probability that the action step is a Slip. Table 1 shows a model trained on data that used the tutor without an agent (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 correlation coefficient of r = 0.460 to the labels, while the detector from the Scooter group data achieved r = 0.481, in each case a moderate degree of correlation [19].

**Table 1.** Carelessness (Contextual Slip) Models for NoScooter and Scooter Groups

| Carelessness (NoScooter) = | Carelessness (Scooter) = |
|---|---|
| -0. 07256 * Answer is right | -0. 11895 * Answer is right |
| -0. 03658 * Action is a bug | -0. 02501 * Action is a bug |
| +0.08997 * Action is a help request | +0. 05535* Input is a choice |
| +0.09944 * Input is a choice | -0. 02876 * Input is a number |
| -0. 03595 * Input is a string | -0. 03772 * Input is a point |
| -0. 02018 * Input is a number | -0.03632 * Input is checkbox or not choice/string/number/point |
| -0. 02805 * Input is a point | +0.04486 * Probability that the student knew the |

skill involved in this action

-0.01662 * Input is checkbox or not choice/string/number/point

+0. 00903 * Probability that the student knew the skill involved in this action

+ 0.00707 * Pknow-direct from log files

- 0.01495 * Not first attempt at skill in this problem

-0.06562 * First transaction on new problem

-0.00573 * Time taken, normalized in terms of SD off average across all students at this step

+0.07257 * Answer not right

+0.00025 * Number of errors the student made on this skill on all problems

-0.00067 * Errors has this student averaged on this skill across problems

+0.00021 * Total time spent on this skill across problems

+0.00532 * Previous 3 actions were on the same cell

-0.00335 * Previous 5 actions were on the same cell

+0.00766 * How many of the previous 8 actions were help requests

-0.00792 * How many of the previous 5 actions were errors

-0.03136 * Has the student made at least 3 errors on this problem step, in this problem

+0.08456

+ 0.07296 * Pknow-direct from log files

+ 0.10466 * Not first attempt at skill in this problem

+0.00434 * Time taken, normalized in terms of SD off average across all students at this step

+0.00249 * Time taken in last three actions, normalized

+0.11895 * Answer not right

-0.00099 * Errors has this student averaged on this skill across problems

-0.00033 * Total time spent on this skill across problems

+ 0.02207 * Previous 3 actions were on the same cell

-0.01615 * Previous 5 actions were on the same cell

-0.01205 * How many of the previous 5 actions were errors

-0.02557 * Has the student made at least 3 errors on this problem step, in this problem

+0.06601

The carelessness detectors passed the tests for model degeneracy in [3, 4]. Within the 127 students' activities, there were a total of 1221 scenarios where the student had three consecutive correct actions per skill, while 419 instances where the student had at least 10 consecutive correct actions. In both cases, the model was not empirically degenerate – the estimate of knowing the skill afterwards did not decrease after these correct actions. The generated carelessness model also passed the theoretical degeneracy test – the maximum of the new contextual P(S) values did not exceed 0.5.

This model was successful at predicting whether the student would perform correctly on the next opportunity to practice the skill, in both the NoScooter and Scooter groups. The contextual-guess-and-slip model achieved prediction of A' = 0.821 for the NoScooter group, and A' = 0.814 for the Scooter group (A' refers to the model's ability to distinguish between a right and wrong answer, with a chance probability of 0.5). Both contextual-guess-and-slip models achieved slightly higher A' values than their baseline BKT counterpart (A' = 0.816 for the NoScooter group, and A' = 0.807 for the Scooter group), although this was not cross-validated. It is worth-noting that with the low number of skills within the Scatterplot Tutor, the potential benefits of the CGS model for reducing over-parameterization are reduced.

In addition to A' values, the goodness of the models were also supported by their Bayesian Information Criterion values for Linear Regression Models [18]. Both models had BIC' values far less than -6 (NoScooter = -414.60, Scooter = -401.21), the cut-off for a model being better than chance [18], making these models better-than-chance indicators of this behavior.

To investigate generalizability, we tested each detector on the opposite data set, i.e. the NoScooter detector was used on the Scooter group dataset and the detector from the Scooter group was used on the NoScooter group dataset. We also tested the detectors with Scatterplot log data from a US school setting [cf. 6]. These interaction logs from the US (described in greater detail in [6]) were gathered from $6^{th}$-$8^{th}$ 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. Table 2 shows the detectors' correlation between the labeled (from Eq. 2 – our CGS equations) and predicted (from our models) slip values in each data set. Within the NoScooter condition data, the detector trained on the Scooter condition data actually performed slightly better (r=0.471) than the detector trained on the NoScooter data (r=0.460). Within the Scooter data, the detector trained on the NoScooter data performed moderately worse (r=0.392) than the detector trained on the Scooter data (r=0.481), although still respectably. These results appear to indicate between mild degradation and no degradation when a carelessness detector is transferred between versions of the tutor with or without an ECA. The asymmetry in transfer between the two environments can be attributed to the fact that the skills and action steps in the NoScooter environment are also present in the Scooter environment, whereas the opposite is not true. When transferred to data from the USA, both of the detectors trained on data from the Philippines performed quite well, performing better in the USA than in the Philippines for all combinations of training and test conditions. This is striking evidence for detector generalizability, when the detectors perform better 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 to no degradation when transferred to different data sets.

**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) | 0.460 | 0.471 |
| Scooter Group Data (PH) | 0.392 | 0.481 |
| NoScooter Group Data (US) | 0.490 | 0.591 |
| Scooter Group Data (US) | 0.537 | 0.605 |

An interesting additional finding was that the Scooter group committed fewer errors compared to the NoScooter group (both PH and US data). Whether or not these errors were careless, it is possible that Scooter's interventions supported future student performance in the tutor

For both test environments, we also examined the values of P(S) according to the model, when certain conditions hold in the data (average predicted P(S) = 0.12 and maximum P(S) = 0.38 across all conditions). One finding is that errors were more likely to be slips when the probability that the student knew the skill before answering was greater than the initial probability $L_O$ for that skill (the 4009 cases in the data where this condition held had an average predicted P(S) = 0.18, compared to the average P(S) = 0.10 where this condition didn't hold). In addition, if a student's successive actions (at least two) for a particular problem step and skill are correct, a subsequent mistake was more likely to be a slip (850 cases where predicted P(S)

increased to an average of 0.20). Slip was even more strongly associated with cases where the student has made very few prior errors on a skill with a high initial knowledge value ($L_O$) (355 cases in the data, average predicted P(S) = 0.27).

## 5 Conclusion

In this paper, we developed detectors of student carelessness within a lesson on scatterplots in a Cognitive Tutor for middle school mathematics, building off prior work in this area [3, 4]. These detectors were tested for robustness when transferred to a different version of the same tutor, and data from schools in a different country. Two carelessness detectors (for the NoScooter condition and the Scooter condition, which incorporated an Embodied Conversational Agent) were created from interaction logs acquired from the tutor usage of Philippine high school students, using a variant of Bayesian Knowledge Tracing, the Contextual Guess and Slip method, which dynamically estimated if an incorrect response was a slip. Our results suggest that these detectors are generalizable and can transfer across tutors with interface differences (i.e. with and without an embodied conversational agent), as well as across different school settings (i.e. Philippine high school and US middle school), increasing potential for automatically intervening in future systems when the students is careless.

**Acknowledgments**
This research was supported by 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", and by the Pittsburgh Science of Learning Center (National Science Foundation) via grant "Toward a Decade of PSLC Research", award number SBE-0836012. 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

[1] Arroyo, I., Cooper, D., Burleson, W., Woolf, B. P., Muldner, K., Christopherson, R., "Emotion Sensors Go To School," In Proceedings of the International Conference on Artificial Intelligence in Education (2009), 17-24.

[2] Baker, R.S.J.d., "Modeling and Understanding Students' Off-Task Behavior in Intelligent Tutoring Systems," *Proceedings of ACM CHI 2007: Computer-Human Interaction* (2007), 1059-1068.

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

[4] Baker, R.S.J.d., Corbett, A.T., Aleven, V., "More Accurate Student Modeling Through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing," *Proceedings of the 9th International Conference on Intelligent Tutoring Systems* (2008), 406-415.

[5] 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., "Contextual Slip and Prediction of Student Performance After Use of an Intelligent Tutor," *Proceedings of the 18th Annual Conference on User Modeling, Adaptation, and Personalization* (2008).

[6] Baker, R. S. d. J., Corbett, A. T., Koedinger, K. R., Evenson, S. E., Roll, I., Wagner, A.Z., Naim, M., Raspat, J., Baker, D. J., & Beck, J., "Adapting to when students game an intelligent tutoring system," *In Proceedings of the 8th International Conference on Intelligent Tutoring Systems* (2006), 392-401.

[7] Baker, R.S.J.d., D'Mello, S.K., Rodrigo, M.M.T., Graesser, A.C., "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 (4) (2010), 223-241.

[8] Baker, R.S.J.d., Gowda, S.M., An Analysis of the Differences in the Frequency of Students' Disengagement in Urban, Rural, and Suburban High Schools. *Proceedings of the 3rd International Conference on Educational Data Mining* (2010), 11-20.

[9] Clements, M.A., "Analysing children's errors on written mathematical tasks," *Educational Studies in Mathematics* (1982), 1–21.

[10] Clements, M.A., "Careless errors made by sixth-grade children on written mathematical tasks," *Journal for Research in Mathematics Education* (1982), 13(2):136–144.

[11] Corbett, A.T., Anderson, J.R., "Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge," *User Modeling and User-Adapted Interaction*, 4 (1995), 253-278.

[12] Craighead, W.E. The Concise Corsini Encyclopedia of Psychology and Behavioral Science, 3rd Edition (2004).

[13] Dix, A., Finlay, J., Abowd, G., Beale, R., Human-Computer Interaction. *Prentice Hall* (1993).

[14] Fogarty, J., Baker, R.S.J.d., Hudson, S.E., Case studies in the use of ROC curve analysis for sensor-based estimates in human computer interaction. *In Proceedings of Graphics Interface* (2005), 129-136.

[15] Levitin, D.J., Ed. Foundations of Cognitive Psychology: Core Readings. MIT Press (2002).

[16] Koedinger, K.R., Anderson, J.R., Hadley, W.H., & Mark, M., "Intelligent tutoring goes to school in the big city," *International Journal of Artificial Intelligence in Education*, 8 (1997), 30-43.

[17] McQuiggan, S. W., Lee, S., and Lester, J. C. Early prediction of student frustration. In A. Paiva, R. Prada, and R. W. Picard (Eds.), Affective Computing and Intelligent Interaction (2007), 698-709.

[18] Rafferty, A. E.. Bayesian model selection in social research. *Sociological Methodology*, 25 (2003), 111-163.

[19] Rosenthal, R., Rosnow, R. L., Essentials of behavioural research: Methods and data analysis, *McGraw-Hill Humanities* (2008).