Affect-Aware Student Models for Robot Tutors

Samuel Spaulding
MIT Media Lab
Cambridge, MA 02139
samuelsp@media.mit.edu

Goren Gordon
Curiosity Lab
Tel-Aviv University
Tel Aviv 6997801, Israel
goren@gorengordon.com

Cynthia Breazeal
MIT Media Lab
Cambridge, MA 02139
cynthiab@media.mit.edu

ABSTRACT

Computational tutoring systems, such as educational software or interactive robots, have the potential for great societal benefit. Such systems track and assess students' knowledge via inferential methods, such as the popular Bayesian Knowledge Tracing (BKT) algorithm. However, these methods do not typically draw on the affective signals that human teachers use to assess knowledge, such as indications of discomfort, engagement, or frustration.

In this paper we present a novel extension to the BKT model that uses affective data, derived autonomously from video records of children playing an interactive story-telling game with a robot, to infer student knowledge of reading skills. We find that, compared to a control group of children who played the game with only a tablet, children who interacted with an embodied social robot generated stronger affective data signals of engagement and enjoyment during the interaction. We then show that incorporating this affective data into model training improves the quality of the learned knowledge inference models.

These results suggest that physically embodied, affect-aware robot tutors can provide more effective and empathic educational experiences for children, and advance both algorithmic and human-centered motivations for further development of systems that tightly integrate affect understanding and complex models of inference with interactive, educational robots.

Keywords

socially assistive robots; educational robots; child-robot interaction; affective computing

1. INTRODUCTION

Intelligent tutoring systems (ITS) are software tutors that attempt to combine the educational benefits of one-on-one tutoring with the scale and ease-of-deployment of software [23]. While ITS research has led to improved outcomes for students, they are often designed as digital workbooks, with the tutoring system providing a sequence of practice problems and offering the student a limited action space (often just the choice of providing an answer or asking for a hint). Recent evidence indicates more interactive learning styles may be better suited for school-age children; we believe the design of computational tutors should reflect these principles.

Previous research indicates that *physically embodied social robots, capable of perceiving and understanding affective signals* can deliver more engaging and empathic educational experiences [13]. In this paper, we add to this increasing body of evidence by establishing social robots' unique capability to engage children in *affect-aware tutoring experiences*.

We conducted an analysis of children's facial expressions from video records of an experiment described in [9] and [10]. In this experiment, children played an interactive story-telling game with a robot. During the game, the robot would assess the child's reading ability by verbally asking them to identify a word written on the screen. The experiment was recorded, including a close camera shot of the child's face, which, along with time-synced records of the child's game actions, forms our dataset.

This analysis was conducted using Affdex, a commercial affect-analysis tool based on FACS and trained on data from over 2.7 million faces in 75 countries [20, 17]. We developed a custom application using the Affdex SDK to analyze 38 children's interactions with a robot, each approximately 12 minutes long.

We compared the affective response, derived via Affdex, of children who interacted with a robot to the affective response of a control group of children who only interacted with a tablet, and show that children who played the story game with a robot generated stronger average emotional expressions over the course of the interaction.

We then used this affective data, in combination with the student's performance data (right/wrong answers), to train a novel, *affective* model for inferring student knowledge. We compare the affective model to a standard Bayesian Knowledge Tracing (BKT) model (trained only on the student performance data) and show that the affective model is better able to generalize to test data from the same population.

2. RELATED WORK

This paper describes work to combine the sophisticated modeling techniques of Intelligent Tutoring Systems with the engaging and empathic abilities of physically embodied, socially interactive robots. Here we give a short review of some relevant work from the ITS and HRI communities.

Intelligent Tutoring Systems (ITSs) refer to a wide variety of computer-based educational tools [23]. The subclass of "affect-aware tutors" [26] are ITSs that explicitly sense, model, and reason about students' affective states. Inspired by psychological theories of emotion and learning, affect-aware tutors seek to foster engagement and learning from data-driven estimates of students' affective states. Previous work has established the validity of affective data captured during educational interactions [11], and efforts to develop affect-aware tutoring systems have culminated in a num-

Appears in: *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016),*
J. Thangarajah, K. Tuyls, C. Jonker, S. Marsella (eds.),
May 9–13, 2016, Singapore.

ber of extensively studied systems, e.g., the Wayang Tutor [1] and Affective Meta-Tutor [24] projects. Yet much of the computational work on affect and modeling in the ITS literature focuses on models to *infer* affect. Typically, once affective states are detected or identified, they trigger simple behavioral rules: a tutor might change its facial expression or offer a supportive comment. These rules are most commonly hardcoded by the developers and remain fixed throughout the deployment, limiting the generality and flexibility of the agent's affective responses.

In parallel, researchers have explored the use of robots as educational tutors. Social robots have been successfully deployed in schools to teach topics as varied as vocabulary [22], chess [13], mathematics [4], and abstract analytic skills [14]. Long-term studies of robots, while less common, have demonstrated that social robots can be an effective tool for improving children's literacy skills [12]. These results suggest that social robots may be particularly well-suited to educational tutoring applications.

Some work has explored physical robots that respond to human affect in education. As with ITSs, however, the robot's responses are typically the result of scripted rules. In work by Szafir and Mutlu [21], students wore an EEG sensor while a robot told a story. The robot responded to perceived student decreases in attention by producing exaggerated gestures, which led to better participant story recall.

Researchers in the LIREC project developed a robotic chess tutoring system that uses more sophisticated affective models to give empathy-based support [13]. The robot models a child's affective state by tracking in-game events in combination with external physical sensors and offers support by mirroring the child's (estimated) affective state.

Unlike previous efforts to unite affect, tutoring, and robots, in this paper, we incorporate affect directly into a robot's *learning* model. While other work has used affect as an input to behavioral rules, this work is the first to evaluate a robotic tutoring system that uses affective data to train models of student knowledge.

3. METHODS FOR COMPUTATIONALLY ASSESSING KNOWLEDGE

One challenge for any tutor, human or computer, is the problem of determining what a student does or does not know. If a student's skill level is too far above or below the level required by the curriculum, they may feel bored or discouraged. The ability to modulate the curriculum to keep students in a state of Flow – in which they are highly engaged in a task at an optimal challenge level – is considered a crucial advantage of one-on-one tutoring [8]. To do so, a tutor must have an accurate estimate of a student's skill level. But how does one assess another's knowledge?

3.1 Bayesian Knowledge Tracing

For computational tutors, Bayesian inference on graphical models is the most popular approach. If the subject domain can be suitably modeled, algorithms for inference can allow computational systems to accurately estimate student skill levels. The most popular and widely used of these methods is known as Bayesian Knowledge Tracing (BKT), a domain-general model used to infer skill mastery from student data. BKT models are widely used in ITS research and can be applied to any educational domain which can be decomposed into different component “skills"[7, 2].

Under the BKT model, a student's mastery of each skill is modeled by a Dynamic Bayesian Network. These skill models are a special case of Hidden Markov Models, which can be in one of two hidden states, “skill mastered" or “skill not-mastered", where the observations are the student's answers to questions requiring knowledge of a particular skill.

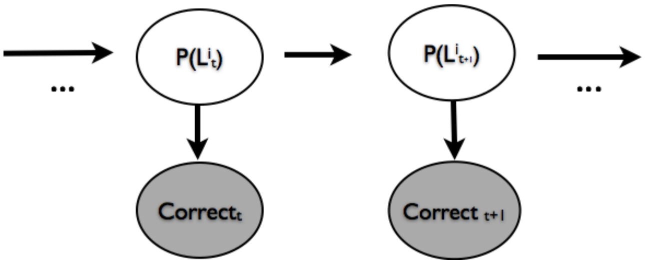

Figure 1: Traditional BKT Hidden Markov Model. Based on observations of a student's correct and incorrect responses, the model can infer the student's hidden knowledge state. The state nodes represent the probability that skill *i* was learned after question t. The observable nodes represent whether the student correctly demonstrated skill *i* when answering question *t*

The BKT model, while popular for its straightforward analysis and ease of implementation, suffers from some limitations. Most significantly for this work, the BKT model relies solely on a student's pattern of correct/incorrect answers to drive inference, while ignoring a wide range of relevant contextual information, such as a student's affective expressions. In spite of BKT's limitations, it is still one of the most popular methods for knowledge assessment and forms the basis of a substantial body of research [2, 27].

3.2 Affective-BKT

In this paper we introduce an affective variant of the BKT model, which we call the Affective-BKT (Aff-BKT) model – a BKT model with additional observation nodes representing features of the student's facial expression measured during an educational interaction with a robot. The Aff-BKT model is an attempt to improve the BKT model, by augmenting it to draw inference from *affective* as well as knowledge-based features. Whereas the standard BKT model has only one observation per timestep (whether the student answered a question correctly or not), the Aff-BKT model is an HMM that incorporates multiple observations per time step. In addition to the observable node corresponding to a correct/incorrect response, our Aff-BKT model includes observable nodes that correspond to whether a child was smiling and whether a child appeared engaged. These affective observation nodes are structured identically to the observation of correct/incorrect answers (see Fig. 1), therefore each additional node requires just two additional parameters per skill model.

While our Affective BKT model does not address all of the limitations of BKT, it serves as a proof-of-concept that affective information can be reliably detected and made useful to computational tutoring systems, and that models that take advantage of affective information can outperform existing techniques.

4. EXPERIMENTAL OVERVIEW

The data used to train and evaluate our models come from video records of a previously published experiment [9, 10]. In this experiment, children (aged 4-8) played a story game with a robot. The interaction was designed around a child-robot-tablet interface, in which the child and robot sit across from each other, with a tablet placed between them as a shared, social context that can be sensed by the robot (Figure 3).

The child and the robot played a custom-designed game called “Storymaker." In the game, graphics of characters (such as animals

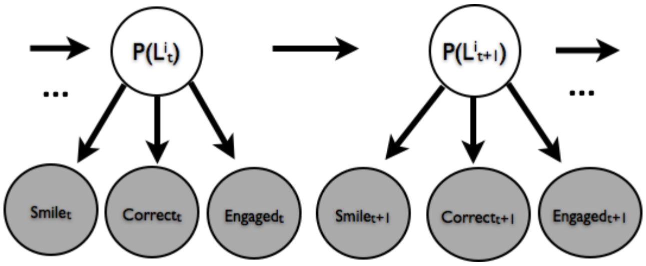

Figure 2: The Affective BKT model, which incorporates affective signals into knowledge state inference.

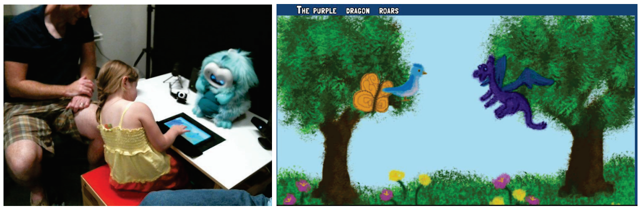

Figure 3: Picture of the interaction setup from which the dataset was collected. Children played a storytelling game with a robot, during which they were periodically prompted to read a word

or objects) float on a background of different scenes (such as a beach or jungle) and can be moved by the child via touchscreen. The robot was framed as a younger peer. It greeted the child by saying "Let's play word games together!" and was introduced by the experimenter as "a young robot who has just learned to speak, and wants to learn to read."

During the game, the child was prompted to move the characters to different regions of the scene. After each 'action' taken by the child, the app procedurally generates a sentence that characterizes this action, which the robot would speak aloud. For example, if the child moves the Dragon character graphic towards the rightmost side of the forest scene (depicting a tree), the game might generate the sentence "Dragon goes to the tree" as part of the story. The spoken sentence would also appear in written form at the top of the tablet.

50% of the time, after the robot had spoken a sentence, the robot would follow by prompting the child to read one of the words in the sentence by saying: "I don't know how to read the word [X], can you show it to me?" The game would then pause until the child tapped on one of the words in the sentence (shown at the top of the tablet). If the child tapped the correct word, the app (via TTS) would read the tapped word, thereby letting the child know whether their answer was correct. If the child tapped an incorrect word, he/she would be prompted to try again by the robot ("I don't think that's right. Can you try again?"). After two incorrect tries, the game would highlight and read the correct word, then continue with the next part of the story. On average, a child experienced 29 of these demonstration opportunities during the interaction.

4.1 Tablet-only Interaction Participants

Section 4 describes the interaction for 25 of the participants. However, there was also a smaller control group of 13 participants that did *not* interact with the robot. In this condition, a white cardboard box was placed over the robot for the duration of the experiment, The children were not told there was a robot underneath the box, only that they would be playing word games on a tablet. Other than the presence of the covering box and the removal of experimenter references to the robot, the experiment proceeded exactly as described above. The robot's story 'speech' still came from the same speakers and the children were still prompted to read words as before; the underlying technology did not change. This condition therefore provides an ideal scenario to examine how children's emotional expressiveness varies with the presence of a social robot during *otherwise identical* interactions, which we discuss in Section 7.1.

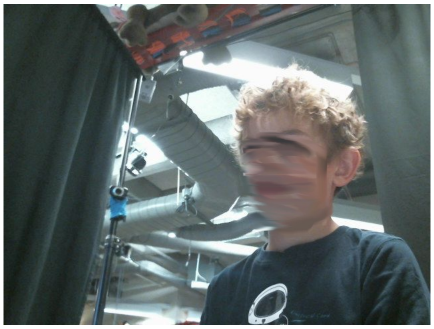

Figure 4: Sample frame of collected footage

4.2 Video QA-interval Dataset

Video of the entire interaction was recorded from a camera, located behind and to the side of the robot, aimed at the child's face (Fig. 4). In addition to video data, the full state of the game, all child actions on the tablet, all robot speech, and all tablet actions (e.g., highlighting the correct word if the child answered incorrectly twice) were recorded and synchronized via ROS to support search, playback, and analysis.

Though complete footage of the interaction was recorded, in this work we are primarily concerned with analyzing children's emotions while they are engaged in educational activity. In addition, the HMM formalism for both Knowledge Tracing models (BKT and Aff-BKT) requires discrete timesteps. We therefore only analyze video footage from specific, educationally relevant, intervals during each interaction called **QA-intervals**.

As described above, after a child moved the story characters, the robot would sometimes ask the child to read a word, denoted [X], by asking the child "I don't know the word [X]. Can you show it to me?" The robot was asking the child to identify word [X] in text on the screen, providing both an opportunity for the child to demonstrate his/her reading ability and an opportunity to collect spontaneous facial expression data during educational activity. We identified the precise times at which the robot *asked* a question and the times at which the child *gave an answer to that question*. We then extracted video footage from 5 seconds before each asking event to 5 seconds after each response event, which defined a **"QA-interval"**.

5. AFFECTIVE BKT: IMPLEMENTATION AND EVALUATION

In this section, we describe the implementation and evaluation of the two Knowledge Tracing models in the context of assessing the students' reading skills. We discuss the skills we chose to model, followed by detailing the structural implementations of each model, how the training data was obtained, and how the parameters for each model were trained from the data.

5.1 Modeling alphabetic principle skills

Learning to read is an acquired skill that requires mastery of many foundational sub-skills before fluent "reading" occurs [25]. In this paper, we focus on modeling a subset of these skills known as "alphabetic principle" skills. The alphabetic principle is the recognition that *written* letters and their combinations correspond to *spoken* sounds in specific and predictable ways. In other words, these skills are the rules and mappings that connect written letters with spoken sounds [6, 16].

In this paper, we model children's literacy progress via **four BKT skill models and four Aff-BKT skill models**. Three models of each type correspond to one of three alphabetic principle skills. In addition, we include one model of each type that tracks the complete ability to read a word exactly. The three alphabetic principle skills tracked were: recognizing the first grapheme of a word, identifying a word of approximately the same length as a spoken word, and recognizing the final grapheme of a word.

The first skill, denoted FIRST-LETTER, requires the child to hear a requested word, decompose its sounds into phonemes, map the phonemes into graphemes, and then select a word that features those graphemes at the beginning of the word. Roughly, it corresponds to the first step of the oft-repeated advice to a young reader - "Sound it out!" We considered a student to have correctly demonstrated this skill if they selected a word that *starts* with the same grapheme as the requested word.

The second skill, denoted LENGTH, requires the child to understand that the length of a written word corresponds to the number of syllables in its spoken instantiation. For this skill, we considered a student to have correctly demonstrated this skill if they selected a word with a length within 1 letter of the requested word's length.

The third skill, denoted LAST-LETTER, requires the child to go through the same process as FIRST-LETTER, applied to the end of the word. This skill poses slightly more of a challenge, as children naturally attempt to read a word from its start. We considered a student to have correctly demonstrated this skill if they selected a word that *ends* with the same grapheme as the requested word.

Lastly, we model the complete, correct reading of a word. This skill, denoted EXACT-CORRECT, requires the child to fully and correctly identify the requested word, representing the complete ability to read the requested word. Because the requested word was spoken aloud by the tablet, homophones were considered equivalent (e.g. if the robot asked for the word 'to', and both 'to' and 'too' were among the possible answers, either would be considered 'exactly correct').

We chose to analyze this set of skills because they are well-suited to the task from which the dataset was derived (see Sec 4), relatively easy to detect computationally (compared to, e.g., analysis of a child's pronunciation), and because they are developmentally appropriate for the age group of the population [19]. Though the alphabetic principle skills are clearly interrelated, BKT skills are most commonly treated independently. In line with our stated research focus (examining whether *affective features* improve Knowledge Tracing models), we assume that each skill is mastered independently of the others.

5.2 Building BKT and Aff-BKT Skill Models

We constructed BKT models for each tracked skill using Kevin Murphy's Bayes Net Toolkit [18], a freely available, open-source library for implementing a wide variety of graphical models. Each BKT model has one hidden node – with two possible states, (Skill Learned or Skill Not-Learned) – and one observation node – with two possible values (Question Correct or Question Incorrect) – per time step. Each BKT model has the structure depicted in Figure 1 and is fully specified by 5 parameters, which set the rate of learning, the probability of forgetting a previously mastered skill, the probability of guessing correctly (if the skill is not mastered), the probability of making a mistake (i.e., answering wrong if the skill is mastered), and the prior probability that the skill is mastered.

As with the BKT models, we constructed an Aff-BKT model for each skill using the Bayes Net Toolkit. Each Aff-BKT model includes two additional observable nodes per time step – $Smile_t$ and $Engaged_t$, depicted in Figure 2. These additional nodes require just two additional parameters each: the probability that a child appears engaged if they have/have not mastered the skill and the probability that a child smiles if they have/have not mastered the skill. Thus, each Aff-BKT model is fully specified by 9 parameters. In the following sections, we describe how we derived the training data from video footage of the interactions, how the parameters for each model were learned from the data, and how we evaluated each model.

5.3 Deriving QA-interval skill data

In Section 4.2, we defined the concept of a *QA-interval*: the educationally relevant portions of the interactions from just before the robot asked for a word to be read to just after the child gave an answer. We implemented string-matching functions to compute whether the word the child provided during each QA-interval represented a correct application of a skill, given the requested word. For example, if a child selected the word **'prince'** when the requested word was **'princess'**, that is a correct application of FIRST-LETTER, but not of EXACT-CORRECT, LENGTH, or LAST-LETTER. For each QA-interval, we applied these functions to the requested and answered words. The end result was, for each QA-interval, four boolean results (Correct or Incorrect) representing whether or not the child's answer represented a correct demonstration of each skill (FIRST-LETTER, LENGTH, LAST-LETTER, EXACT-CORRECT).

5.4 Deriving QA-interval affective data

5.4.1 Affdex: a tool for autonomous affect detection

Often, researchers studying human affect use human coders to manually review video footage of interactions and provide labels from a set of pre-determined affective states. Sometimes coders will have training in facial expression coding, but in many cases, the video coders are either the experimenters themselves, undergraduate research assistants, or online workers (e.g., workers on Amazon Mechanical Turk). Inter-coder reliability ratings can help verify the accuracy of labeling, but all of these methods are ultimately subject to human judgement.

Our research goals are to develop social robot tutors that are robust enough to be widely deployed for long-term interactions and intelligent enough to act autonomously. To achieve these goals, robots cannot rely solely on human annotation to understand user affect. Methods for *unobtrusively* and *autonomously* sensing affect are rapidly improving; in line with our vision, we rely solely on autonomous affect detection and labeling in this work.

To analyze children's emotional expressions without relying on human coders, we used the Affdex SDK, a commercial tool marketed by Affectiva, Inc. to enable development of affect-aware mobile applications. Affdex uses state-of-the-art face detection and analysis algorithms to extract estimates of four physical facial expression features **(Smile, BrowFurrow, BrowRaise, and LipDepress)** and two hidden affective features **(Valence and Engagement)** from video or images of faces. For each of these six metrics, Affdex produces an estimate of the strength of that affective indicator, normalized to [0, 100], excepting Valence, which is normalized to [-100, 100].

Affdex uses data collected from over 2.7 million faces in authentic scenarios, labeled by FACS-trained coders and rigorously validated by leading specialists in affective computing [20, 17]. We used the Affdex SDK to analyze sequences of time-stamped images extracted from video. Affdex analyzes each frame (providing data at a much finer level of granularity than is typically afforded by human coders) and each frame result can be precisely linked back to the rest of the data through its timestamp.

5.4.2 *Analyzing affective data*

We sampled still frames from video footage of each QA-interval (each interaction contained 29 QA-intervals, on average) at approximately 20fps. We wrote a simple app using the Affdex SDK that analyzed all frames from each QA-interval. The output of the analysis was, for each frame, either a single measured value for each of the 6 metrics, or a null measurement (indicating that no data could be extracted from the frame). Each frame result was then written to file, and manually transferred to a secure data storage location for subsequent analysis.

Autonomously sensed affective data is difficult to capture and interpret. Unsurprisingly, the initial measurements from Affdex were highly variable, even on a frame-by-frame basis. Frame results would alternate rapidly between consistent, modest values (indicative of correct processing and analysis) and null or extreme results (e.g., minimum or maximum values for all metrics). To smooth these discontinuities, we applied a median filter to the data, operating over a sliding window of 20 frames, (see Fig. 5). The end result of this median smoothing process is a set of data points with significantly fewer spikes or rapid increases and decreases. However, when such increases or decreases are present, they are more likely due to a *bona fide* affective signal, rather than noise.

To derive data for the Smile/Engagement features of the Affective-BKT model, for each QA-interval we calculated the mean value of the Engagement and Smile metrics, both of which have been used as input to behavioral rules in prior affective-aware tutoring work [5]. We assigned each QA-interval a discrete, boolean label of Smile/No Smile and Enaged/Not Engaged via a mean value threshold of 30. The end result of this process was, for each QA-interval, two sets of boolean results (Smile/No Smile and Engaged/Not Engaged). Note that, unlike the skill correctness data, the affective data does not change with the skill being modeled.

5.5 Training BKT and Aff-BKT models from QA-interval data

5.5.1 *Expectation Maximization: learning model parameters from observed data*

We trained the parameter values for each skill model via Expectation Maximization (EM). Expectation Maximization (EM) is a general technique for estimating a set of parameters from data. Under the assumption that some set of parameters, θ, generated a set of data, D, Expectation Maximization tries to find the Maximum Likelihood Estimate (MLE) of the parameters, given the data: $max_{\theta} Pr(\theta|D)$. For the BKT models, the data used to train each skill model is the set of skill-correctness data for the corresponding skill (described in Sec. 5.3). For the Aff-BKT skill models, the training data is the skill-correctness data *and* the affective data. EM converges when the difference in likelihood between iterations falls below some threshold parameter, ε. In this work, we let $\varepsilon = 10^{-5}$. All models converged in <15 iterations. Each BKT model took approximately 20 minutes to train; each Aff-BKT model took approximately 40 minutes to train.

5.5.2 *Initial Conditions*

EM is a deterministic algorithm, and therefore the final learned model is somewhat sensitive to initial conditions. We initially set the prior mastery parameter at .5, the learning rate parameter at .2, the slip and guess parameters at .25 (the actual chance of guessing correctly ranges from $11-33\%$ depending on the skill and QA-interval), and, as is common in short-term BKT models, the forgetting rate parameter at 0. During EM, these parameters are tuned to maximize the likelihood of the parameter set, given the data. The end result of training is a set of learned parameters (i.e., the learned model), which we denote θ_{aff} for an Aff-BKT model and $\theta_{bkt}$for a traditional BKT model.

6. BAYESIAN MODEL SELECTION

The main research question of this paper is whether incorporating affective information into the training process improves BKT models. Knowledge Tracing is difficult precisely because a student's knowledge state is not directly observable. Because we have no ground-truth data, it is infeasible to use traditional supervised learning benchmarks (e.g., precision and recall or F-score), to evaluate the Aff-BKT and BKT models.

To overcome this limitation, we evaluate our models from the perspective of Bayesian Model Selection. Bayesian Model Selection is the general problem of determining which of several possible models should be preferred, given some data. Approaches to model selection vary based on the space of possible models and datasets, but typically rely on comparing the likelihoods of different models.

Many standard model selection metrics (e.g., Bayesian or Akaike Information Criterion) are used to compare models with different *structures*, trained on the same *data*. The BKT and Affective-BKT models differ not only in their structure but also in the data they model and are trained from; therefore, these techniques are not appropriate for this case.

To properly address our stated research focus, we construct a new Aff-BKT model, denoted $\hat{\theta}_{aff}$, which has the same *structure* as θ_{bkt}, but is trained from the Aff-BKT *data*. This model is the subset of the parameters that the BKT and Aff-BKT models share. Specifically, we derive the $\hat{\theta}_{aff}$ model by taking the subset of parameters from θ_{aff} that are also present in θ_{bkt} (described in Section 5.2) after training.

By doing so, we reduce the difficulty of model comparison to a more tractable situation. $\hat{\theta}_{aff}$ and θ_{bkt} share the same *structure*, but only $\hat{\theta}_{aff}$ was trained with affective *data*. This enables us to conduct a more straightforward analysis of the models' respective likelihoods: leave-one-out cross-validation (LOOCV), in which all but one participants' data is used to train the model, then the likelihood of each trained model is evaluated, with respect to the held-out participant's data.

LOOCV occurs in two phases: during the "training" phase, a single participant's data is held-out and the two models are trained on the remaining data via EM. Then, during the "testing phase", the probability of the held-out data, under the trained model is calculated (separately, for each model) as an estimate of the model's overall performance. The probability of the held-out data is the likelihood of the model, and is used to measure model 'fit' – that is, how well the model explains the observed data. Models with higher likelihood and, hence, better fit are preferred.

During a full LOOCV evaluation, this process is repeated for each of the participants, such that for a complete analysis of n participants, n different training sessions occur (each differing slightly, due to variation in training data), and n different test data points (the probabilities of all n participant's held-out data or, equivalently, n different model likelihoods) are collected per model.

Data from 38 participants was used. Thus, we repeated the process 38 times, holding out one participant each fold, and computed

Table 1: T-test results by metric, comparing differences in mean interaction value across Robot and Tablet conditions

Metric	T-test statistic	p-value
Smile	2.23	***0.037***
BrowFurrow	0.898	0.378
BrowRaise	2.68	***0.011***
LipDepress	0.328	0.746
Valence	0.844	0.405
Engagement	2.59	***0.014***

38 model likelihoods per skill model. Figure 6 shows the complete evaluation pipeline - from training data, to model construction, to model evaluation. Full results by skill and model type are presented in Figures 8 and 9. Aggregate statistics for each skill and model type are presented in Table 2.

7. RESULTS AND DISCUSSION

7.1 Children are more emotionally expressive when interacting with a social robot

To determine a participant's overall level of emotional expressiveness, we combined the affective data (for each metric) from a participant's QA-intervals and calculated the *mean value* for that metric. In other words, we calculated the average value, over the participant's interaction, of the median-filtered Affdex measurements. Because we are interested in *overall levels* of emotional expression and because valence, unlike other metrics, ranges from [-100, 100] rather than [0,100], we used the mean of the *absolute value* of the valence data. This captures the idea that highly positive and highly negative measurements both represent 'strong' emotional expressions and are evaluated as such in our analysis. Figure 5 depicts the complete data analysis workflow, from video footage to mean interaction metric value.

For each metric, we obtained mean interaction values from 25 participants that interacted with a robot, and 13 control group participants that did not interact with a robot. The means and standard error of the mean for each metric and population (Robot vs. Tablet) are shown in Figure 7.

For each metric, we first established the normality of the mean interaction values via a Shapiro-Wilk test. Then, for each metric, we conducted a Student's T-test to determine whether participants who played the game with a physical robot present were, on average, more emotionally expressive.

The test results (Table 1) show that Affdex recorded higher average emotional metric values from children in the robot condition, compared to the children in the tablet-only condition. These differences were statistically significant for 3 of the 6 metrics: **Smile, BrowRaise, and Engagement**. While the other metric differences did not reach significance, in all metrics the average measured value was higher in the robot condition.

These results agree with previous qualitative research, suggesting that children tend to smile and are highly engaged by social robots [3] and extends that research by confirming *quantitatively* and *autonomously* (i.e., without human judgement) that children who interacted with a robot during an educational task generated higher average emotional expression readings than children who did not see the robot. Previous research that has shown that the mere *physical presence* of robots can lead to increased learning gains in tutoring interactions [15]. Our work suggests that *a physical robot's capacity to modulate student affective response* may be one mechanism by which physical presence improves learning gains.

Skill	Model	Mean LL ±*SD*
Exact-Correct	A-BKT	-14.923 ± 9.01
Exact-Correct	BKT	-19.79 ± 10.386
First-Letter	A-BKT	-13.652 ± 10.091
First-Letter	BKT	-17.090 ± 10.386
Length	A-BKT	-13.6160 ± 9.596
Length	BKT	-14.5992 ± 9.513
Last-Letter	A-BKT	-14.644 ± 9.493
Last-Letter	BKT	-18.222 ± 9.98

Table 2: Leave-one-out cross-fold validation log-likelihood metrics of Aff-BKT and BKT models

7.2 Affective-BKT models fit test data better than standard BKT models

Figure 8 shows the the log-likelihood of each BKT and Aff-BKT skill model, evaluated on every participant. For three of the four skills (Exact-Correct, First-Letter, and Last-Letter) the Aff-BKT model error is lower for nearly all participant test data. For Length, the Aff-BKT model error is generally lower, but there were some cases in which the BKT model had better fit. We suspect our approach to modeling the Length skill may have been too coarse to accurately model skill mastery.

Overall, however, it is clear that the model trained with affective data fits the held-out skill data much better than the BKT model. Table 2 reinforces these results: every Aff-BKT skill model has higher mean likelihood than the traditional BKT model for the same skill. Figure 9 shows the statistical significance of these results, presented as a decrease in model error, rather than an increase in 'fit'. Following [27], we use the *negative* log-likelihood as an error function for model evaluation (for ease of presentation). This definition of error follows naturally from the interpretation of model likelihood as a measure of 'fit' (see Sec. 6). Figure 9 shows the mean error (that is, *average negative log-likelihood*) of each skill model.

These results answer our research question – "Can we build models that use affective data to improve the performance of traditional Knowledge Assessment models?" – affirmatively. Interestingly, this conclusion is reached by evaluating a BKT model and an Aff-BKT model that is *structurally identical* to the traditional BKT model and hence does not *explicitly* model affect. Rather, the influence of the affective data manifests itself in the trained values of the traditional BKT parameters. In other words, by including the additional parameters, structure, and data to support affect-awareness, we find that the Aff-BKT model learns better fitting parameters *even for parts of the model that have nothing to do with affect*! Without further study, the underlying reasons for this cannot be conclusively determined. However, we hypothesize that the affective data helps "explain away" some of the observed variance in Correct/Incorrect answers. That is, the BKT model tries to fit all five of its parameters to explain the *full* variance in the skill-correctness data. However, modern theories of affect and learning suggest that the two are deeply related, hence variance in the skill-correctness data can likely be explained in part by affective factors. The additional affective parameters of the Aff-BKT model can adjust to the variance due to affective variables (e.g., distraction, confusion), leaving the remaining parameters (common to the traditional BKT model) free to better fit the variance due to *knowledge-based* factors (e.g., genuine skill mastery).

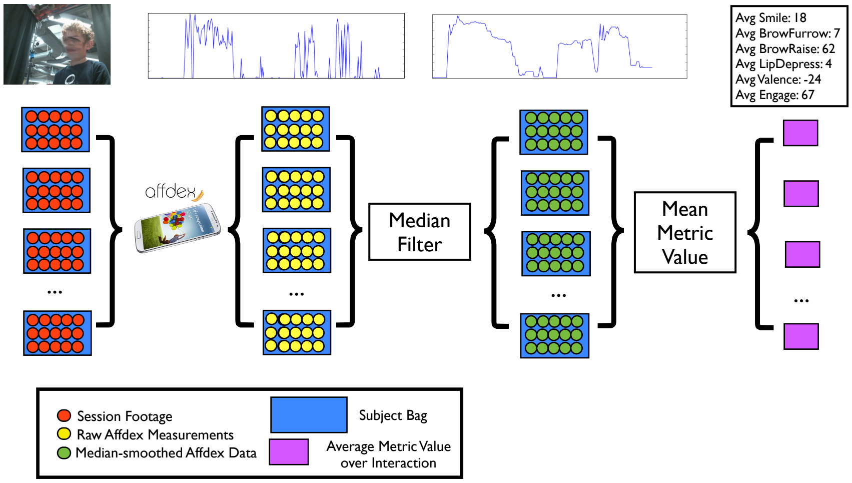

Figure 5: Data pipeline for assessing children's average level of emotional expression. Frames are sampled from video footage at 20fps, then passed to Affdex. The Affdex measurements are then passed through a median filter for smoothing. For each metric, the smoothed data points for each participant are averaged to produce a mean metric value across the entire interaction.

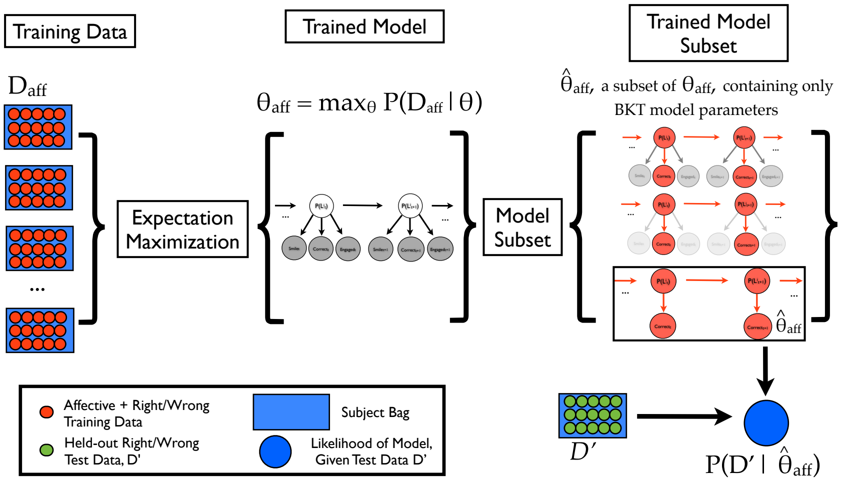
Training Data
Trained Model
Trained Model Subset
D_{aff}
$\theta_{aff} = \max_\theta P(D_{aff} \mid \theta)$
$\hat{\theta}_{aff}$, a subset of θ_{aff}, containing only BKT model parameters
Expectation Maximization
Model Subset
$\hat{\theta}_{aff}$
...
Affective + Right/Wrong Training Data
Held-out Right/Wrong Test Data, D'
Subject Bag
Likelihood of Model, Given Test Data D'
D'
$P(D' \mid \hat{\theta}_{aff})$

Figure 6: A single fold of the full training and evaluation pipeline for Aff-BKT skill models. The models are trained from the right/wrong *and* the affective QA-interval data via EM, with a single participant's data "held-out" for testing, resulting in a set of learned model parameters. We use a subset of the learned parameters, $\hat{\theta}_{aff}$, structured identically to the BKT parameter set, and evaluate both in the same way: comparing the likelihood of the two parameter sets, under the held-out participant's skill-correctness data. This process is repeated once for each participant, thus the final result is a set of 38 'likelihood estimate' data points for each (Aff-BKT and BKT) skill model.

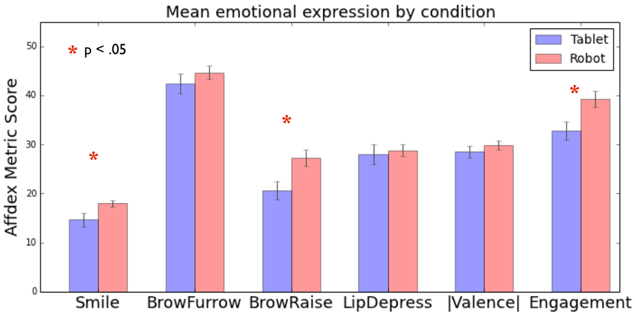

Figure 7: Average emotional expression value, by condition. Error bars represent standard error of the mean.

8. CONTRIBUTIONS AND CONCLUSIONS

In this work, we have shown that children are more emotionally expressive when engaged in an interactive educational task with a social robot than when engaged in an *identical* task with a tablet alone. Previous research has shown that the mere physical presence of social robots can alter important interaction dynamics: our work adds to that body of research by identifying an important aspect of that phenomenon (children's increased emotional expressivity) and demonstrating one way in which the next generation of computational tutors can leverage interactive learning styles for algorithmic improvements.

We have introduced a novel *affective* approach to Knowledge Tracing and demonstrated that children's emotional expression data can be successfully integrated into inferential models for assessing knowledge. These affective models outperform traditional approaches to Knowledge Tracing, demonstrating the utility of sensing affective data and constructing tutoring models to make use that data, suggesting that embodied social robots may be a more beneficial medium for developing affect-aware computational tutors.

Researchers in psychology and cognitive science are coming to understand that affect, far from being an 'irrational' influence, is

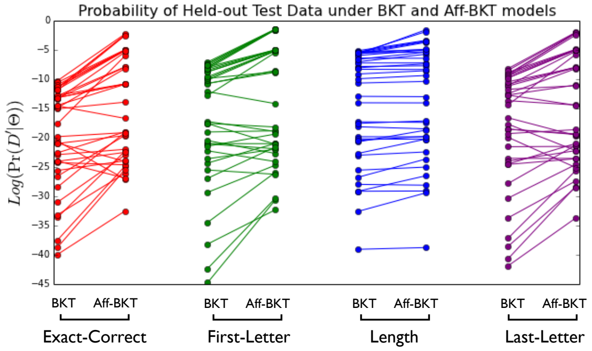

Figure 8: Probability of Test Data under BKT and Aff-BKT models, color-coded by skill. Within each skill, data points to the left represent the log-likelihood (model fit) of the BKT model evaluated on one test participant's data. Each point is connected to the log-likelihood of the Aff-BKT model evaluated on the same participant's data.

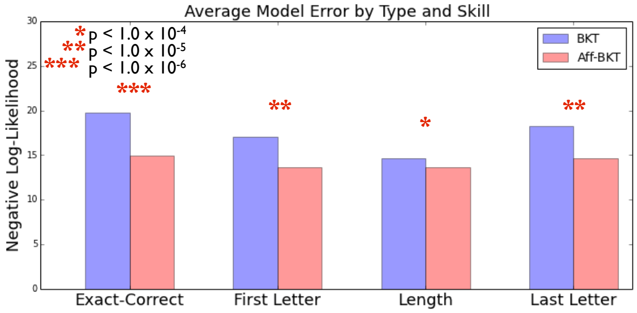

Figure 9: Mean error estimates of BKT and Aff-BKT models by skill. Significance was calculated via a one-sample T-test on the *difference* between model error estimates, evaluated on the same test participant data.

crucial to everyday decision-making. This understanding is beginning to influence the design of agents and algorithms that model or simulate intelligent behavior. While a relatively recent effort, many researchers are now pursuing efforts to integrate affect understanding into intelligent systems. Our work shows that physical robots may have advantages over software-only systems in sensing and using affective data, by demonstrating social robots' ability to induce higher degrees of emotional expressivity and showing that integrating emotionally expressive data into inference can improve the performance of student Knowledge Tracing models. Encouraged by our results, we believe this research will ultimately lead to more effective, intelligent, and natural interactions with technology.

9. ACKNOWLEDGEMENT

This research was supported by the National Science Foundation (NSF) under Grants CCF-1138986 and an NSF GRF under Grant No. 1122374. Any opinions, findings and conclusions, or recommendations expressed in this paper are those of the authors and do not represent the views of the NSF.

REFERENCES

[1] I. Arroyo, C. Beal, T. Murray, R. Walles, and B. P. Woolf. Web-based intelligent multimedia tutoring for high stakes achievement tests. In *Intelligent Tutoring Systems*, pages 468–477. Springer, 2004.

[2] R. S. Baker, A. T. Corbett, and V. Aleven. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In *Intelligent Tutoring Systems*, pages 406–415. Springer, 2008.

[3] T. Belpaeme, P. E. Baxter, R. Read, R. Wood, H. Cuayáhuitl, B. Kiefer, S. Racioppa, I. Kruijff-Korbayová, G. Athanasopoulos, V. Enescu, et al. Multimodal child-robot interaction: Building social bonds. *Journal of Human-Robot Interaction*, 1(2):33–53, 2012.

[4] L. N. Brown and A. M. Howard. The positive effects of verbal encouragement in mathematics education using a social robot. In *Integrated STEM Education Conference (ISEC), 2014 IEEE*, pages 1–5. IEEE, 2014.

[5] W. Burleson. *Affective learning companions: strategies for empathetic agents with real-time multimodal affective sensing to foster meta-cognitive and meta-affective approaches to learning, motivation, and perseverance*. PhD thesis, Massachusetts Institute of Technology, 2006.

[6] B. Byrne and R. Fielding-Barnsley. Phonemic awareness and letter knowledge in the child's acquisition of the alphabetic principle. *Journal of Educational Psychology*, 81(3):313, 1989.

[7] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.

[8] M. Csikszentmihalyi. Flow and education. *NAMTA journal*, 22(2):2–35, 1997.

[9] G. Gordon and C. Breazeal. Bayesian active learning-based robot tutor for children?s word-reading skills. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI-15)*, 2015.

[10] G. Gordon, C. Breazeal, and S. Engel. Can children catch curiosity from a social robot? In *Human-Robot Interaction (HRI), 2015 10th ACM/IEEE International Conference on*, 2015.

[11] J. M. Harley, F. Bouchet, M. S. Hussain, R. Azevedo, and R. Calvo. A multi-componential analysis of emotions during complex learning with an intelligent multi-agent system. *Computers in Human Behavior*, 48:615–625, 2015.

[12] J. Kory and C. Breazeal. Storytelling with robots: Learning companions for preschool children's language development. In *Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on*, pages 643–648. IEEE, 2014.

[13] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva. Empathic robots for long-term interaction. *International Journal of Social Robotics*, 6(3):329–341, 2014.

[14] D. Leyzberg, S. Spaulding, and B. Scassellati. Personalizing robot tutors to individuals' learning differences. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 423–430. ACM, 2014.

[15] D. Leyzberg, S. Spaulding, M. Toneva, and B. Scassellati. The physical presence of a robot tutor increases cognitive learning gains. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society*, 2012.

[16] I. Y. Liberman, D. Shankweiler, and A. M. Liberman. The alphabetic principle and learning to read. 1989.

[17] D. McDuff, R. Kaliouby, and R. W. Picard. Crowdsourcing facial responses to online videos. *Affective Computing, IEEE Transactions on*, 3(4):456–468.

[18] K. Murphy et al. The bayes net toolbox for matlab. *Computing science and statistics*, 33(2):1024–1034, 2001.

[19] U. of Oregon Center for Teaching and Learning. Big ideas in beginning reading: Alphabetic principle, April 2015. [Online; Accessed April 22, 2015].

[20] R. Swinton and R. El Kaliouby. Measuring emotions through a mobile device across borders, ages, genders and more.

[21] D. Szafir and B. Mutlu. Pay attention!: designing adaptive agents that monitor and improve user engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 11–20. ACM, 2012.

[22] F. Tanaka and S. Matsuzoe. Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction*, 1(1), 2012.

[23] K. VanLehn. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221, 2011.

[24] K. VanLehn, W. Burleson, M.-E. C. Echeagaray, R. Christopherson, J. G. Sanchez, J. Hastings, Y. H. Pontet, and L. Zhang. The affective meta-tutoring project: How to motivate students to use effective meta-cognitive strategies. In *19th International Conference on Computers in Education, Chiang Mai, Thailand*, 2011.

[25] M. Wolf and C. J. Stoodley. *Proust and the squid: The story and science of the reading brain*. Icon Cambridge, 2008.

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

[27] M. V. Yudelson, K. R. Koedinger, and G. J. Gordon. Individualized bayesian knowledge tracing models. In *Artificial Intelligence in Education*, pages 171–180. Springer, 2013.