# Using Data from Real and Simulated Learners to Evaluate Adaptive Tutoring Systems

José P. González-Brenes[1], Yun Huang[2]

[1] Pearson School Research & Innovation Network, Philadelphia, PA, USA
jose.gonzalez-brenes@pearson.com
[2] Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA
yuh43@pitt.edu

**Abstract.** Classification evaluation metrics are often used to evaluate adaptive tutoring systems— programs that teach and adapt to humans. Unfortunately, evidence suggests that existing convention for evaluating tutoring systems may lead to suboptimal decisions. In a companion paper, we propose Teal, a new framework to evaluate adaptive tutoring. In this paper we propose an alternative formulation of Teal using simulated learners. The main contribution of this novel formulation is that it enables approximate inference of Teal, which may useful on the cases that Teal becomes computationally intractable. We believe that this alternative formulation is simpler, and we hope it helps as a bridge between the student modeling and simulated learners community.

## 1 Introduction

Adaptive systems teach and adapt to humans and improve education by optimizing the subset of *items* presented to students, according to their historical performance [3], and on features extracted from their activities [6]. In this context, items are questions, or tasks that can be graded individually. Adaptive tutoring may be evaluated with randomized control trials. For example, in a seminal study [3] that focused on earlier adaptive tutors, a controlled trial measured the time students spent on tutoring, and their performance on post-tests. The study reported that the adaptive tutoring system enabled significantly faster teaching, while students maintained the same or better performance on post-tests

Unfortunately, controlled trials can become extremely expensive and time consuming to conduct: they require institutional review board approvals, experimental design by an expert, recruiting and often payment of enough participants to achieve statistical power, and data analysis. Automatic evaluation metrics improve the engineering process because they enable less expensive and faster comparisons between alternative systems.

The adaptive tutoring community has tacitly adopted conventions for evaluating tutoring systems [4]. Researchers often evaluate their models with classification evaluation metrics that assess the *student model* component of the tutoring system— student models are the subsystems that forecast whether a learner will answer the next item correctly. However, automatic evaluation metrics are intended to measure an outcome of the end user. For example, the PARADISE [9] metric used in spoken dialogue systems correlates to user satisfaction scores. We are not aware of evidence that supports that classification metrics correlate with learning outcomes; yet there is a growing body of evidence [2, 5] that suggests serious problems with them. For example, classification metrics ignore that an adaptive system may not help learners— which could happen with a student model with a flat or decreasing learning curve [1, 8]. A decreasing learning curve implies that student performance decreases with practice; this curve is usually interpreted as a modeling problem, because it operationalizes that learners are better off with no teaching.

We study a novel formulation of the Theoretical Evaluation of Adaptive Learning Systems (Teal) [5] evaluation metric. The importance of evaluation metrics is that they help practitioners and researchers quantify the extent that a system helps learners.

## 2 Theoretical Evaluation of Adaptive Learning Systems

In this section, we just briefly summarize Teal and do not compare it with a related method called ExpOppNeed [7]. Teal assumes the adaptive tutoring system is built using a single-skill Knowledge Tracing Family model [3, 6]. Knowledge Tracing uses a Hidden Markov Model (HMM) per skill to model the student's knowledge as latent variables. It models whether a student applies a practice opportunity of a skill correctly. The latent variables are used to model the latent student proficiency, which is often modeled with a binary variable to indicated mastery of the skill.

To use Teal on data collected from students, we first train a model using an algorithm from the Knowledge Tracing family, then we use the learned parameters to calculate the effort and outcome for each skill.

- Effort: Quantifies how much practice the adaptive tutor gives to students. In this paper we focus on counting the number of items assigned to students but, alternatively, amount of time could be considered.
- Outcome: Quantifies the performance of students after adaptive tutoring. For simplicity, we operationalize performance as the percentage of items that students are able to solve after tutoring. We assume that the performance on solving items is aligned to the long-term interest of learners.

Algorithm 1 describes our novel formulation. Teal calculates the expected number of practice that an adaptive tutor gives to students. We assume that the tutor stops teaching a skill once the student is very likely to answer the next item correctly according to a model from the Knowledge Tracing Family [6]. The adaptive tutor teaches an additional item if two conditions hold: (i) it is likely that the student will get the next item wrong— in other words, the probability of answering correctly the next item is below a threshold $\tau$; and (ii) the tutor has not decided to stop instruction already.

The inputs of Teal are:

- Real student performance data from $m$ students practicing a skill. Data from each student is encoded into a sequence of binary observations of whether the student was able to apply correctly the skill at different points in time.
- A threshold $\tau \in \{0 \ldots 1\}$ that indicates when to stop tutoring. We operationalize this threshold as the target probability that the student will apply the skill correctly.
- A parameter $T$ that indicates the number of practice opportunities each of the simulated students will practice the skill.

**Algorithm 1** Teal algorithm for models with one skill per item

```
Require: real student data $\mathbf{y}^{(1)} \ldots \mathbf{y}^{(m)}$, threshold $\tau$, # of simulated time steps $T$
function TEAL
    $\theta \leftarrow$ Knowledge_Tracing($\mathbf{y}^{(1)} \ldots \mathbf{y}^{(m)}$)
    $e \leftarrow \{ \}$
    $s \leftarrow \{ \}$
    for $\hat{\mathbf{y}} \in$ get_simulated_student($\theta, T$) do:
        $e \leftarrow$ calculate_effort($\hat{\mathbf{y}}, \theta, \tau$)
        if $e < T$ then
            $s \leftarrow$ calculate_score($\hat{\mathbf{y}}$, e)
        else
            $s \leftarrow$ imputed_value
    return mean($e$), mean($s$)
```

Teal learns a Knowledge Tracing model from the data collected from real students interacting with a tutor. Our new formulation uses simulated learners sampled from the Knowledge Tracing parameters. This enables us to decide how many simulated students to generate. Our original formulation required $2^m$ sequences to be generated, which can quickly become computationally intractable. If an approximate solution is acceptable, our novel formulation allows more efficient calculations of Teal. Teal quantifies the effort and outcomes of students in adaptive tutoring. Even though measuring effort and outcomes is not novel by itself, Teal's contribution is measuring both without a randomized trial. Teal quantifies effort as how much practice the tutor gives. For this, we count the number of items assigned to students. For a single simulated student, this is:

$$\text{calculate\_effort}(y_1, \ldots, y_T, \theta, \tau) \equiv \arg\min_t p(y_t | y_1 \ldots y_{t-1}, \theta) > \tau \tag{1}$$

The threshold $\tau$ implies a trade-off between student effort and scores and responds to external expectations from the social context. Teal operationalizes the outcome as the performance of students after adaptive tutoring as the percentage of items that students are able to solve after tutoring:

$$\text{calculate\_score}(y_1, \ldots, y_T, e) \equiv \sum_{t=e} \frac{\delta(\mathbf{y}_t, \text{correct})}{T - e} \tag{2}$$

Here, $\delta(\cdot, \cdot)$ is the Kronecker function that returns 1 iff its arguments are equal.

## 3 Discussion

Simulation enables us to measure effort and outcome for a large population of students. Previously, we required Teal to be computed exhaustively on all student outcomes possibilities. We relax the prohibitively expensive requirement of calculating all student outcome combinations. Our contribution is that Teal can be calculated with a simulated dataset size that is large yet tractable.

## References

1. R. Baker, A. Corbett, and V. Aleven. More Accurate Student Modeling through Contextual Estimation of Slip and Guess Probabilities in Bayesian Knowledge Tracing. In B. Woolf, E. Aïmeur, R. Nkambou, and S. Lajoie, editors, *Intelligent Tutoring Systems*, volume 5091 of *Lecture Notes in Computer Science*, pages 406–415. Springer Berlin / Heidelberg, 2008.
2. J. Beck and X. Xiong. Limits to accuracy: how well can we do at student modeling? In S. K. D'Mello, R. A. Calvo, and A. Olney, editors, *Proceedings of the 6th International Conference on Educational Data Mining, Memphis, Tennessee, USA, July 6-9, 2013*, pages 4–11. International Educational Data Mining Society, 2013.
3. A. Corbett and J. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1995.
4. A. Dhanani, S. Y. Lee, P. Phothilimthana, and Z. Pardos. A comparison of error metrics for learning model parameters in bayesian knowledge tracing. Technical Report UCB/EECS-2014-131, EECS Department, University of California, Berkeley, May 2014.
5. González-Brenes and Y. José P., Huang. Your model is predictive— but is it useful? theoretical and empirical considerations of a new paradigm for adaptive tutoring evaluation. In J. G. Boticario, O. C. Santos, C. Romero, and M. Pechenizkiy, editors, *Proceedings of the 8th International Conference on Educational Data Mining*, Madrid, Spain, 2015.
6. J. P. González-Brenes, Y. Huang, and P. Brusilovsky. General Features in Knowledge Tracing: Applications to Multiple Subskills, Temporal Item Response Theory, and Expert Knowledge. In M. Mavrikis and B. M. McLaren, editors, *Proceedings of the 7th International Conference on Educational Data Mining*, London, UK, 2014.
7. J. I. Lee and E. Brunskill. The impact on individualizing student models on necessary practice opportunities. In K. Yacef, O. R. Zaïane, A. Hershkovitz, M. Yudelson, and J. C. Stamper, editors, *Proceedings of the 5th International Conference on Educational Data Mining*, pages 118–125, Chania, Greece, 2012.
8. D. Rai, Y. Gong, and J. E. Beck. Using dirichlet priors to improve model parameter plausibility. In T. Barnes, M. Desmarais, C. Romero, and S. Ventura, editors, *Proceedings of the 2nd International Conference on Educational Data Mining*, Cordoba, Spain, 2009.
9. M. Walker, C. Kamm, and D. Litman. Towards developing general models of usability with PARADISE. *Natural Language Engineering*, 6(3):363–377, 2001.