PREDICTING ROBUST LEARNING WITH THE VISUAL FORM OF THE MOMENT-BY-MOMENT LEARNING CURVE

Ryan S.J.d. Baker[1], Arnon Hershkovitz[1], Lisa M. Rossi[2], Adam B. Goldstein[2], Sujith M. Gowda[2]

[1]Department of Human Development, Teachers College Columbia University, 525 W. 120th Street, New York, NY 10027

[2]Department of Social Science and Policy Studies, Worcester Polytechnic Institute, 100 Institute Road, Worcester, MA 01609

Corresponding Author: Ryan S.J.d. Baker, Department of Human Development, 453 Grace Dodge Hall, Teachers College Columbia University, 525 W. 120th Street, New York, NY 10027, baker2@exchange.tc.columbia.edu, +1-212-678-8329

Abstract

We present a new method for analyzing a student's learning over time, for a specific skill: analysis of the graph of the student's moment-by-moment learning over time. Moment-by-moment learning is calculated using a data-mined model which assesses the probability that a student learned a skill or concept at a specific time during learning (Baker, Goldstein, & Heffernan, 2010, 2011). Two coders labeled data from students who used an intelligent tutoring system for college genetics, in terms of seven forms that the moment-by-moment learning curve can take. These labels are correlated to test data on the robustness of students' learning. We find that different visual forms are correlated with very different learning outcomes. This work suggests that analysis of moment-by-moment learning curves may be able to shed light on the implications of students' different patterns of learning over time.

Keywords: Moment-by-moment learning, learning curve, intelligent tutoring system, educational data mining

Introduction

Over recent decades, there has been increasing evidence that very fine-grained analyses of how individual students' performance shifts over time can support deeper scientific understanding of how student knowledge, cognition, and reasoning change over time and based on specific experiences (cf. Siegler & Crowley, 1991). However, current paradigms for studying these types of changes in student knowledge and skill have often had to choose between studying learning over relatively brief periods, for small numbers of subjects, or with less richness than is possible. In this paper, we present a novel tool for tracing students' learning over time in intelligent tutoring systems, the Moment-by-Moment Learning Curve, which represents a step towards addressing these limitations, supporting quick and relatively rich analysis of the conditions under which learning occurs. We demonstrate this tool's potential through an analysis which shows that a combination of visual analysis of the Moment-by-Moment Learning Curve and data mining can produce a model that infers how robust student learning is, enabling prediction of which students retain their knowledge over time and/or are prepared for future learning.

Tracing Learning Over Time

Research on individual learning over time has been conducted in several fashions. We can group the majority of the methods used into three paradigms: 1) Microgenetic methods, 2) Qualitative video analyses, and 3) Visual analysis. Microgenetic methods have been a very popular method for studying individual students' learning over time, for over two decades (cf. Siegler & Jenkins, 1989). Implementing a microgenetic approach involves collecting repeated observations and measurements of a student's behavior and learning during a given learning process, ideally during a period of substantial change in the learner. These observations and measurements are intensively analyzed with the aim of determining the processes responsible for driving both qualitative and quantitative change in the student’s state or knowledge (Siegler & Crowley, 1991). In Siegler and Jenkins’ work, for instance, microgenetic analysis revealed that students of varying school years solve different types of subtraction problems using different strategies, but that specific types of mathematics problems can trigger the development of more sophisticated strategies. Kuhn and colleagues (1992, 2008) applied the microgenetic method in their research to study theory revision and strategy change made by students engaging in self-directed exploration during problem solving, and to study the development of argument skills in a computer-supported environment. This research suggests that in addition to breaking down processes of change, the microgenetic method also sheds light on the fundamental nature of the skills being learned. One of the main advantages of microgenetic methods is their applicability across domains and in different learning environments. However, a major limitation of microgenetic methods has been the difficulty in scaling to large amounts of data or large numbers of students; current microgenetic methods require considerable time to study individual students.

A second commonly used paradigm is qualitative video analysis to study changes in student reasoning over time, using data from video collected over a period of time, be it a single session or a full semester. As with microgenetic methods, this approach can be challenging to extend to large amounts of data or to large numbers of students. For example, within this paradigm, Cobb (1999) analyzed students’ mathematical learning as it occurred in social interactions within a classroom environment over a 10-week period. The social atmosphere promoted qualitative argumentation in the classrooms, prompting students to switch from an additive mode of thinking to a multiplicative mode of thinking which improved their overall mathematical reasoning skills over time. Additionally, Lehrer, Schauble, Strom, and Pligge (2001) demonstrate the importance of modeling in learning math and science concepts like weight, volume, and density, which can often be difficult for students. They found that having students create graphs to represent weight as a function of volume for different types of materials helped students differentiate density from weight, and helped the students conceptualize the relationship between these three properties of matter.

A third category comprises tools and techniques for tracing students' learning and change of knowledge, using visual representations. The most straightforward example is the use of learning curves, which display changes in student performance for a specific knowledge component (skill or concept) over time (cf. Estes, 1950; Mazur & Hastie, 1978). The statistical parameters underlying the learning curve quantify the learning that is occurring. In the context of computer-based learning environments, learning curves may be easily generated based on log data of student performance within the software (e.g., Martin et al., 2011; Mathan & Koedinger, 2005). Log data can also be used for generating other visual representations of learning-related variables over time, such as affect (e.g., D'Mello & Graesser, 2011). An alternate visual representation of learning over time is Hershkovitz and Nachmias's (2009) learnograms, which are visual representations of multiple learning process-related variables over time. Whereas learning curves are often used to represent populations of students, learnograms are typically used to represent the performance or behavior of a single student over time. For instance, Hershkovitz and Nachmias (2009) use learnograms to show how a given student switches between multiple learning activities. They also use learnograms to study how the intensity with which students engage in specific learning activities varies over time. Given the focus of learnograms on individual students, they have been used to study individual students' learning trajectories, a type of usage that is less common with learning curves, which are typically used to study overall patterns across a group of learners or differences between groups of learners.

A key limitation to learnograms, the microgentic method, and qualitative video analysis is the difficulty in scaling to large amounts of data or large numbers of students. The patterns seen in learnograms can be relatively complicated to visually analyze, and learnograms must be analyzed one-by-one for each student and construct, reducing their benefit for analyzing complex patterns across students or groups of students. The microgenetic method is slower still, requiring considerable time to study individual students. Using qualitative analysis of video to study change in students requires time consuming coding methods prior to analysis even beginning. As such, these methods each are limited in the scope of problems which can be studied. Specifically, current paradigms for studying learning over time have been limited in their usefulness for studying individual student differences and changes, across substantial numbers of students, and over long periods of time; these methods are too time-consuming to easily use to conduct analyses at scale. Learning curves are, by contrast, more scalable, but are not easy to use to study individual differences.

Another key limitation of learning curves and learnograms is their focus on performance/behavior. Despite the fact that their names include the word "learning," these representations do not so much show learning as they show performance or behavior – correctness, time taken, and use of specific learning resources at specific times. As such, learning must be inferred in a relatively indirect fashion. In a learning curve, it is clear that a group of students who makes half as many errors after a set of practice opportunity has learned, but it is not clear when or under what conditions specific students' learning occurred. This limitation is less present in the case of slower to apply microgenetic methods and qualitative video analysis, where specific moments of learning are often identified (e.g., Siegler & Jenkins, 1989; Cobb, 1999).

The Moment-by-Moment Learning Curve

However, a recent advance in student modeling creates the possibility of studying student learning over time, in a fashion that is scalable but also allows in-depth analysis of individual learners. This work, *analysis of the moment-by-moment learning curve,* creates visual representations of learning as it occurs moment-by-moment, in a three-step process. In this method, the probability that learning occurred at a specific moment is inferred, based on the probability that the student has learned the step up to that point, and the probability of their future actions given the probability that they learned the step at that moment.

The first step is to use a Bayesian Knowledge Tracing (BKT) model (Corbett & Anderson, 1995) to calculate the probability that the student knows a specific skill at a specific time, based on the student's history of success on problems or problem steps involving that skill. Skills are assigned to each problem or problem step (either through knowledge engineering or data mining – cf. Cen et al., 2006), and the BKT model is updated every time the student responds to a problem, based on the correctness of the response, allowing for an aggregate estimate of student knowledge over time.

In the second step, a recent model proposed by Baker, Goldstein, and Heffernan (2010, 2011) is used to build on this estimation to infer the probability that a student learned a skill or a concept at a specific step during the problem-solving process. That is to say, instead of assessing the probability that a skill is known at time N, the model assesses the probability that the skill was learned between time N-1 and time N. As Bayesian Knowledge Tracing infers student knowledge from correctness (cf. Corbett & Anderson, 1995), this model's operational definition of learning is when a student's performance shifts from being (mostly) incorrect to (mostly) correct (taking the probability of guessing and slipping into account).

It is worth briefly considering the difference between the approach used here, and prior work that attempted to identify when student performance transitions to completely correct. Research using the "mean trial of last error" paradigm studied how long students took to begin producing consistently correct answers, using this measure to compare between learning conditions (e.g., Bower, 1961; Bower & Trabasso, 1963). This paradigm, while useful in the simple concept learning paradigms it was used for, has a key limitation for complex problem-solving domains, such as typically seen in education. In those domains, it is common to see students "slipping" and producing incorrect answers, on occasion, even after learning a skill. Restricting analysis to identifying the point where performance becomes error-free may miss earlier points where performance improves substantially (but not to the point of complete perfection). By contrast, studying the probability of transition from "mostly incorrect" to "mostly correct" allows for inference about the degree of improvement over time.

The calculation of this model is discussed in detail later in the paper, but integrates across information about past performance, current performance, and future performance, using a combination of Bayesian formulas and data mining (cf. Romero & Ventura, 2007; Baker & Yacef, 2009), to compute a probability that a skill or concept was learned at a specific time. Derivatives of this model have been shown to be effective at predicting several learning constructs, including predicting a student's eventual knowledge (Baker, Goldstein, & Heffernan, 2010, 2011), and predicting tests of preparation for future learning of entirely different skills

(Baker, Gowda, & Corbett, 2011). It is worth emphasizing that these two models infer learning, and the degree of learning at a specific moment, based on probabilities. The probabilities used to calculate learning at a specific moment are themselves estimated using the full data. For instance, the probability that a student achieved a correct answer by guessing (rather than by knowing the skill) is estimated during the initial calculation of the BKT model.

In the third step, this model's estimates are graphed over time for a specific student and skill, producing a moment-by-moment learning curve. This graph represents a specific student's moment-by-moment learning for a given skill. Moment-by-moment learning curves have several advantages over traditional approaches to studying learning over time. First, these graphs visualize the measurement of learning (rather than performance) at a fine-grain level. This facilitates consideration of what features of an action (whether by the student, or an intervention directed towards the student) are associated with greater momentary learning, a statistically challenging question in previous paradigms (cf. Beck, 2006; Beck & Mostow, 2008). Second, a separate moment-by-moment learning curve can be created for each individual student, on a specific skill. In past works on traditional learning curves of performance, it has been argued that curves computed across students have different properties than curves computed for individual students (Anderson & Tweney, 1997; Heathcote et al., 2000). However, key visual analyses conducted on learning curves such as looking for unexpectedly difficult items indicating a poor skill-item mapping (cf. Corbett & Anderson, 1995) have not been feasible to conduct on graphs of individual students (for instance, graphs of correctness over time appear as 1s and 0s for single students). Moment-by-moment learning curves, by comparison, are easily interpretable for individual students.

The first analyses of moment-by-moment learning curves focused on a specific metric, termed spikiness (Baker, Goldstein, & Heffernan, 2010, 2011), intended to show to what extent the graph has peaks that are far above other points, in order to infer if a skill is learned suddenly in a "eureka" moment (Lindstrom & Gulz, 2008) or more gradually (Newell & Rosenbloom, 1981; Heathcote et al., 2000). As the graph visualizes the probability that learning occurred at each opportunity to use a skill, a peak in the graph corresponds to a sudden improvement in performance. To assess this, the following metric was used: maximum moment-by-moment learning in a graph, divided by the average moment-by-moment learning in the graph.

This spikiness metric was found to predict both final student knowledge in a tutoring system (Baker, Goldstein, & Heffernan, 2010, 2011), and tests of preparation for future learning of entirely different skills (Baker, Gowda, & Corbett, 2011). However, in looking more closely at the data, we found that the metric of spikiness obscures considerable information. For example, the two (theoretical) graphs in Figure 1 have the same level of spikiness, but clearly represent very different patterns of learning. The graph on the left has a single spike. This represents a student who learned the current skill at a specific point in time. Though their earlier learning may prepare them for the "eureka" moment shown, and they may continue to strengthen their memory of the skill after the spike (cf. Cen et al., 2007), the spike represents a qualitative shift from not knowing the skill to knowing the skill, which likely corresponds to a qualitative shift from failing to demonstrate the skill to successfully demonstrating it. The graph on the right has two spikes. This represents a student whose performance had two substantial jumps – perhaps from never demonstrating the skill, to sometimes demonstrating the skill, to almost always demonstrating it. Two jumps may represent the transition from partially-correct knowledge to completely-correct knowledge, perhaps through addition or removal of a constraint on a cognitive rule (cf. Singley & Anderson, 1989). Another possibility is that double-spikes are seen for multi-faceted knowledge components (e.g., multiple skills are being treated as a single skill, as in Corbett & Anderson, 1995), and that each spike represents the student learning these sub-skills.

(Place Figure 1 approximately here)

In qualitatively analyzing graphs of the moment-by-moment learning curves, we have discovered several prominent visual patterns. However, we do not yet know if it matters which pattern a student demonstrates. There is evidence that spikiness generally predicts student knowledge and preparation for future learning, but this does not necessarily indicate whether differences in the form that spikiness takes matter. Hence, as a step towards understanding whether differences in the visual form of the moment-to-moment learning curve matter, and what the important differences might be, this paper studies whether different visual forms are associated with different learning outcomes. We leverage existing data from research on robust learning in college genetics (Corbett et al., 2011), including log data from use of an intelligent tutoring system (Koedinger & Corbett, 2006), and data from tests of students' problem-solving skill, ability to transfer knowledge to new situations (cf. Singley & Anderson, 1989; Fong & Nisbett, 1991), preparation for future learning (Bransford & Schwartz, 1999), and retention of skill in later weeks (Schmidt & Bjork, 1992). Hence, we investigate the degree to which differences in moment-by-moment learning curves can predict these indicators of robust learning.

To study this issue, we label a set of moment-by-moment learning curves in terms of their visual forms, and then correlate the frequency of each visual form to learning outcome data. We find that some visual forms of the moment-by-moment learning curves are associated with substantially better learning than other visual forms, and we conclude by considering the implications of this result.

Data

Learning activity. The data set used in the analyses presented here was drawn from the Genetics Cognitive Tutor (Corbett et al., 2010), a learning system that aims to support students in developing abductive reasoning and problem-solving skills in the domain of college Genetics. Within this Cognitive Tutor, students work individually with the tutoring software. This tutor consists of 19 modules that support problem solving across a wide range of topics in genetics, including pedigree analysis and carrier probabilities of pedigrees, gene interaction and epistasis, three-factor cross, gene regulation, and equilibrium and departures of population. Various subsets of the 19 modules have been piloted at 15 universities in North America. This study focuses on a tutor module that employs a gene mapping technique called *three-factor cross*, in which students infer the order of three genes on a chromosome based on offspring phenotypes, as described in (Baker, Corbett, et al., 2010). A screenshot of this module is given in Figure 2. In this study, 72 undergraduates enrolled in genetics or in introductory biology courses at Carnegie Mellon University used the three-factor cross module. The students used the software as a homework assignment, as part of their regular course assignments, but the study was conducted in a laboratory setting, to support administration of all tests. The students engaged in Cognitive Tutor-supported activities for one hour in each of two sessions. All students completed standard three-factor cross problems in both sessions. During the first session, some students were assigned to complete other cognitive-tutor activities designed to support deeper understanding; however, no differences were found between conditions for any learning measure (cf. Corbett et al., 2011), so in this analysis we collapse across the conditions and focus solely on student behavior and learning within the standard problem-solving activities undertaken after completion of the other activities. The 72 students completed a total of 22,885 problem solving attempts across 10,966 problem steps in the tutor.

(Place Figure 2 approximately here)

Pre/post-tests. A pre-test of student ability to solve problems in the tutor was given prior to usage. Post-tests, given by paper-and-pencil, consisted of four activities (cf. Baker, Corbett, et al., 2010): a straightforward problem-solving post-test, a transfer test, a test of preparation for future learning (PFL), and a delayed retention test administered one week after the student completed the software. The straightforward problem-solving post-test and retention test consisted of the same types of items seen in the tutor; three forms were developed, and each form served as a pre-test for 1/3 of the students, a post-test for 1/3 of the students, and a retention test for the remaining 1/3 of students. The transfer test included two problems intended to tap students' understanding of the underlying processes. The first was a three-factor cross problem that could not be solved with the standard solution method and required students to improvise an alternative method. The second problem asked students to extend their reasoning to four genes. It provided a sequence of four genes on a chromosome and asked students to reason about the crossovers that must have occurred in different offspring groups. The PFL test consisted of 2½ pages of instruction on the reasoning needed for an analogous, but more complex, four-factor cross gene mapping task, followed by a single four-factor cross problem for students to solve.

Students demonstrated successful learning in this tutor, with an average pre-test performance of 0.33 (SD = 0.2), an average post-test performance of 0.83 (SD = 0.19), and an average PFL performance of 0.89 (SD = 0.15). The various post-tests were only moderately correlated with one another, as shown in Table 1. The two most correlated tests were the transfer and problem-solving tests (r = 0.59) and the two least correlated tests were the retention and PFL tests (r = 0.33). Hence, we will analyze the four tests separately.

(Place Table 1 approximately here)

Creating the Moment-by-Moment Learning Curve

As noted earlier, the moment-by-moment learning curve is composed of predictions made by the moment-by-moment learning model (Baker, Goldstein, & Heffernan, 2010, 2011). Within this section, we discuss how the model predicts the probability P(***J***) that a student has learned a specific knowledge component at a specific problem step (***J*** stands for for "Just learned"). This model is developed using the following approach:

First, training labels of the probability that a student learned a knowledge component at a specific problem step are generated, using a combination of predictions of current student knowledge from Bayesian Knowledge Tracing and data on future correctness, integrated using Bayes' Theorem. In essence, we use evidence from both the past and future to assess the probability that learning occurred at a specific time.

Then, a data-mined model is built using a broad set of features calculated from these labels. Most importantly, this model includes absolutely no data from the future. The result is a model that can be used either at run time or retrospectively, to assess the probability that a KC is learned at each practice opportunity. Refining the original training labels with data mining in this fashion improves the model's predictions of individual actions. Whereas the original labels only use a limited degree of data, the data mined labels boost these labels with additional data features, and information about the level of P(***J***) associated with those features, across the whole data set. Hence, limitations or noise in the original labels can be reduced by data mining. This data mined model has been previously used to study the relationship between the spikiness of graphs – at an aggregate level – and learning, and has been shown to predict student final knowledge (cf. Baker, Goldstein, & Heffernan, 2010, 2011) and preparation for future learning (Baker, Gowda, & Corbett, 2011). In the following sub-sections, we detail the labeling process, the construction of the features set, and the building of the machine-learned model.

Labeling Process

The first step of the process of creating the moment-by-moment learning model is to label each problem step ***N*** in the data set (i.e., the ***N***th opportunity for the given student to use the given KC) with the probability that the student learned the KC at that time, to serve as inputs for machine learning. Our specific working definition of "learning at step ***N***" is learning the KC between the instant after the student enters their first answer for step ***N***, and the instant that the student enters their first answer for step ***N+1***.

We label step ***N*** using information about the probability that the student knew the KC before answering on step ***N*** (from Bayesian Knowledge Tracing) and information about performance on the two following steps (***N+1***, ***N+2***). Using data from future actions gives information about the true probability that the student learned the KC during the actions at step ***N***. For instance, if the student probably did not know the KC at step ***N*** (according to Bayesian

Knowledge Tracing), but the first attempts at steps ***N+1*** and ***N+2*** are correct, it is relatively likely that the student learned the KC at step ***N***. Correspondingly, if the first attempts to answer steps ***N+1*** and ***N+2*** are incorrect, it is relatively unlikely that the student learned the KC at step ***N.*** Note that these calculations are based on estimated probabilities from the BKT model, which are themselves estimated using the entire data set, rather than being absolute decisions based on heuristics.

We can assess the probability that the student learned the KC at step ***N,*** given information about the actions at steps ***N+1*** and ***N+2*** (which we term ***A*****+1+2**), as:

$$\mathrm{P}(\boldsymbol{J}) = \mathrm{P}(\boldsymbol{\sim L_n \wedge T} \mid \boldsymbol{A_{+1+2}})$$

We can find P(***J***)'s value with a function using Bayes' Rule:

$$P(\sim L_n \wedge T \mid A_{+1+2}) = \frac{P(A_{+1+2} \mid \sim L_n \wedge T) * P(\sim L_n \wedge T)}{P(A_{+1+2})}$$

The base probability P(***~Ln*** ^ ***T***) can be computed fairly simply, using the student's current value for P(~***Ln***) from Bayesian Knowledge Tracing, and the Bayesian Knowledge Tracing model's value of P(***T***) for the current KC:

$$P(\sim \boldsymbol{L_n} \wedge \boldsymbol{T}) = P(\sim \boldsymbol{L_n}) P(\boldsymbol{T})$$

The probability of the actions at time ***N+1*** and ***N***+2, P($\boldsymbol{A}_{+1+2}$), is computed as a function of the probability of the actions given each possible case (the KC was already known, P(***Ln***), the KC was unknown but was just learned, P(***~Ln*** ^ ***T***), or the KC was unknown and was not learned, P(***~Ln*** ^~***T***), and the contingent probabilities of each of these cases.

$$P(\boldsymbol{A}_{+1+2}) = P(\boldsymbol{A}_{+1+2} \mid \boldsymbol{L_n}) P(\boldsymbol{L_n}) + P(\boldsymbol{A}_{+1+2} \mid \sim \boldsymbol{L_n} \wedge \boldsymbol{T}) P(\sim \boldsymbol{L_n} \wedge \boldsymbol{T}) + P(\boldsymbol{A}_{+1+2} \mid \sim \boldsymbol{L_n} \wedge \sim \boldsymbol{T}) P(\sim \boldsymbol{L_n} \wedge \sim \boldsymbol{T})$$

The probability of the actions at time ***N+1*** and ***N***+2, in each of these three cases, is a function of the Bayesian Knowledge Tracing model's probabilities for guessing (***G***), slipping (***S***), and learning the KC (***T***). In order to calculate the probability of each possible case of estimated student knowledge, we must consider all four potential scenarios of performance at actions ***N+1*** and **N+*2***. In the formulas below, correct answers are written **C** and non-correct answers (e.g., errors or help requests) are written **~C**. The possible scenarios are: correct/correct (**C**, **C**); correct/wrong (**C**, **~C**); wrong/correct (**~C**, **C**); and wrong/wrong (**~C**, **~C**):

$$\begin{array}{ll}
P(\boldsymbol{A_{+1+2}} = \boldsymbol{C}, \boldsymbol{C} |\, \boldsymbol{L_n}) = P(\sim\boldsymbol{S})^2 & P(\boldsymbol{A_{+1+2}} = \boldsymbol{C}, \sim\boldsymbol{C} |\, \boldsymbol{L_n}) = P(\boldsymbol{S})P(\sim\boldsymbol{S}) \\
P(\boldsymbol{A_{+1+2}} = \sim\boldsymbol{C}, \boldsymbol{C} |\, \boldsymbol{L_n}) = P(\boldsymbol{S})P(\sim\boldsymbol{S}) & P(\boldsymbol{A_{+1+2}} = \sim\boldsymbol{C}, \sim\boldsymbol{C} |\, \boldsymbol{L_n}) = P(\boldsymbol{S})^2 \\
P(\boldsymbol{A_{+1+2}} = \boldsymbol{C}, \boldsymbol{C} |\, \sim\boldsymbol{L_n}\hat{}\boldsymbol{T}) = P(\sim\boldsymbol{S})^2 & P(\boldsymbol{A_{+1+2}} = \boldsymbol{C}, \sim\boldsymbol{C} |\, \sim\boldsymbol{L_n}\hat{}\boldsymbol{T}) = P(\boldsymbol{S})P(\sim\boldsymbol{S}) \\
P(\boldsymbol{A_{+1+2}} = \sim\boldsymbol{C}, \boldsymbol{C} |\, \sim\boldsymbol{L_n}\hat{}\boldsymbol{T}) = P(\boldsymbol{S})P(\sim\boldsymbol{S}) & P(\boldsymbol{A_{+1+2}} = \sim\boldsymbol{C}, \sim\boldsymbol{C} |\, \sim\boldsymbol{L_n}\hat{}\boldsymbol{T}) = P(\boldsymbol{S})^2
\end{array}$$

$$P(\boldsymbol{A_{+1+2}} = \boldsymbol{C}, \boldsymbol{C} |\, \sim\boldsymbol{L_n}\hat{}\sim\boldsymbol{T}) = P(\boldsymbol{G})P(\sim\boldsymbol{T})P(\boldsymbol{G}) + P(\boldsymbol{G})P(\boldsymbol{T})P(\sim\boldsymbol{S})$$

$$P(\boldsymbol{A_{+1+2}} = \boldsymbol{C}, \sim\boldsymbol{C} |\, \sim\boldsymbol{L_n}\hat{}\sim\boldsymbol{T}) = P(\boldsymbol{G})P(\sim\boldsymbol{T})P(\sim\boldsymbol{G}) + P(\boldsymbol{G})P(\boldsymbol{T})P(\boldsymbol{S})$$

$$P(\boldsymbol{A_{+1+2}} = \sim\boldsymbol{C}, \boldsymbol{C} |\, \sim\boldsymbol{L_n}\hat{}\sim\boldsymbol{T}) = P(\sim\boldsymbol{G})P(\sim\boldsymbol{T})P(\boldsymbol{G}) + P(\sim\boldsymbol{G})P(\boldsymbol{T})P(\sim\boldsymbol{S})$$

$$P(\boldsymbol{A_{+1+2}} = \sim\boldsymbol{C}, \sim\boldsymbol{C} |\, \sim\boldsymbol{L_n}\hat{}\sim\boldsymbol{T}) = P(\sim\boldsymbol{G})P(\sim\boldsymbol{T})P(\sim\boldsymbol{G}) + P(\sim\boldsymbol{G})P(\boldsymbol{T})P(\boldsymbol{S})$$

Once each action is labeled with estimates of the probability P(***J***) that the student learned the KC at that time, we use these labels to create machine-learned models that can accurately predict P(***J***) at run time. The original labels of P(***J***) were developed using future knowledge, but the machine-learned models predict P(***J***) using only data about the action itself (no future data).

Features

In order to predict the training labels of P(***J***) created in the previous step, we distill a set of features that can be used as predictors. These features are quantitative (or binary) descriptors of key aspects of each problem step that have a reasonable potential to be statistically associated with the construct of interest: whether learning occurred at a specific moment. These features are then used within machine learning (discussed in the next section).

For each problem step (in this learning system, a problem consists of one or more steps, each of which pertains to a specific skill), we used a set of features describing the first action on problem step ***N***. The list consisted of 23 features previously distilled to use in the development of contextual models of guessing and slipping (cf. Baker, Corbett, & Aleven, 2008). These features had in turn been used in prior work to develop automated detectors of off-task behavior (Baker, 2007) and gaming the system (Baker et al., 2008). The actual features selected for incorporation into the final models are given in a following section, in Table 2. The list of features inputted into the machine learning algorithm was:

- Details about the action:
    - The tutoring software's assessment of the action – was the action correct, incorrect and indicating a known bug (procedural misconception), incorrect but not indicating a known bug, or a help request?
    - The type of interface widget involved in the action – was the student choosing from a pull-down menu, typing in a string, typing in a number, plotting a point, or selecting a checkbox?
    - Was this the student's first attempt to answer (or obtain help) on this problem step?
- Measurements of time:
    - How many seconds the action took.
    - 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, or 5, 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. (two variables).
    - How many seconds the student spent on each opportunity to practice the primary skill involved in this action, averaged across problems.
- Previous interaction:
    - The total number of times the student has gotten this specific problem step wrong, across all problems (includes multiple attempts within one problem).
    - What percentage of past problems the student made errors on this problem step
    - The number of times the student asked for help or made errors at this skill, including previous problems.
    - How many of the last 3 actions involved this problem step.
    - How many of the last 5 actions involved this problem step.
    - How many times the student asked for help in the last 8 actions.
    - How many errors the student made in the last 5 actions.
- Knowledge Assessment:
    - The tutor's assessment, before and after the action, of the probability that the student knows the skill involved in this action, derived using Bayesian Knowledge Tracing (two variables)
- Other measurements:
    - Total practice opportunities on this KC so far.

Machine Learning

Given the labels and the model features for each student action within the tutor (Baker, Goldstein, & Heffernan, 2011), we conducted linear regression within RapidMiner (Mierswa et al., 2006) to develop models that predict P(***J***). This resulted in a set of numerical predictions of P(***J***), one for each problem step that a student completed. In each case, M5’ feature selection (Hall, 2000) was used to determine which features were incorporated into the models. Linear regression with M5’ feature selection creates regression trees, a tree of linear regression models, and then conducts linear regression on the set of features used in the tree. The machine learned models generated for each system (including all features in the final models) are listed below in Table 2.

To validate the generalizability of our models, we checked our results with six-fold cross-validation at the student level (i.e., detectors are trained on five groups of students and tested on a sixth group of students). By cross-validating at this level, we increase confidence that detectors will be accurate for new groups of students.

The goodness of the models was validated using the Pearson correlation coefficient between the training labels of P(***J***) for each step, and the values predicted for P(***J***) for the same step by the machine-learned models. As both set of values are quantitative, and there is a one-to-one mapping between training labels and predicted values, linear correlation is a reasonable metric.

The P(***J***) model achieved a solid correlation of 0.676 to the training labels under six-fold student-level cross-validation. The best-fitting model, trained across all data, is given in Table 2.

(Place Table 2 approximately here)

Although the degree of correlation was acceptable, one curious aspect of this model is that it tended to underestimate values of P(***J***), particularly those that were relatively high in the original labels. The difference between the model values of P(***J***) and the original label is highly correlated to the original label, with a correlation of 0.749 in the Cognitive Tutor. Hence, the predicted values of P(***J***) for training labels with high values remained higher than the predicted values of P(***J***) for training labels with lower values (hence the model's reasonable correlation to the labels). However, the predicted values of P(***J***) for training labels with high values were lower, in absolute terms, than the original training labels for those data points. This problem could be addressed by weighting the (rarer) high values more heavily during model-fitting, although this approach would likely reduce overall correlation. Another possible solution would be to fit the data using a logarithmic (or other) function that scales upwards more effectively than a linear function; as will be seen later, the differences between maximum and minimum spikiness are large enough that non-linear regression may be more appropriate than our current approach. Nevertheless, within the current model it is likely to be more straightforward to interpret differences in P(***J***) than absolute values. As such, the curve visual forms we analyze are solely in terms of relative differences, rather than absolutes.

Graph Replays of Visual Forms of the Moment-by-Moment Learning Curve

In order to study the implications of different visual forms of the moment-by-moment learning curve, we created what we term "graph replays," graphs of the moment-by-moment learning curve ready to be labeled by human coders. This work builds off past work in pretty-printing log files as text, termed text replays (cf. Baker et al., 2006; Sao Pedro et al., 2010, in press), in the analysis of learning curves (Martin et al., 2011; Mathan & Koedinger, 2005), and in visualizing student behavior over time (Hershkovitz & Nachmias, 2009). Like text replays, there is automated support for human coders for sampling sub-sets of the data for analysis, looking at a visual representation of each sub-set of data, quickly labeling it, and automatically collating the data.

The choice to leverage human judgment in identifying curve visual forms, rather than pre-defining specific functions, was chosen for several reasons. The initial idea of analyzing curve visual form came during research on spikiness and learning (Baker, Goldstein, & Heffernan, 2011) when it was observed that many graphs had multiple spikes. Defining what a multiple spike is mathematically is a challenging process, whereas it is easy for a human being to identify a graph with multiple spikes (as we discuss in the results section, good inter-rater reliability was achieved). Furthermore, several distinct visual forms were noticed when qualitatively examining data. Defining mathematical functions for each is considerably more challenging than identifying visual forms, making this a good situation to leverage human pattern-recognition skills (cf. Henderson, 1999).

In order to facilitate visual inspection, Java code was written by the fourth author. The program presents the user with a visual display of how a specific student’s moment-by-moment learning changes over time, for a specific skill, across multiple opportunities to demonstrate that skill. The user can then click a set of check boxes in order to identify which visual features the graph possesses and lacks. The user was given the option to click multiple check boxes, because many of the visual graph features we identified were not mutually exclusive (for instance, closely-occurring multiple spikes are not inconsistent with also having a plateau).

Each of the 72 students’ entire activity on each skill was used as the basis of a graph replay, with one exception: if a student completed their work on a skill in under 5 problems, no graph replay was generated. This exception was due to the difficulty of assessing visual form with four data points or fewer.

A set of seven potential replay tags were chosen:

- Single spike – One action with a significantly higher P(***J***) than the rest of the student's responses; this pattern might indicate a sudden moment of learning or "eureka moment," where a difficult skill is suddenly understood.
- Close multi-spike – Several closely clustered actions with significantly higher P(***J***) values than neighboring actions; this pattern suggests a consecutive set of learning events.
- Separated multi-spike – Several actions, not clustered together closely, with significantly higher P(***J***) values than neighboring actions; this pattern might correspond to a multi-phase, more gradual learning.
- Plateau – Three or more sequential actions that have significantly higher P(***J***) values than the rest of the student's behavior; this patterns implies a period of continual learning.
- Constant – No substantial changes in P(***J***) value across the entire replay; this pattern indicates very stable continual learning (which may indicate high continual learning, or no learning at all).
- Immediate peak – First action has a high P(***J***) value, followed by an even higher value, which then immediately falls to low values for the rest of the replay; this pattern might imply that the student quickly learned the skill once starting the tutor activity.
- Immediate drop – First action has a high P(***J***) value, which then immediately falls to low values for the rest of the replay; this pattern might tell us that the student already knew the skill before starting to use the tutor.

In addition, it was possible for coders to label a graph with an unusual data feature as "undefined." Graphs corresponding to each of these tags are shown in Figure 3.

(Place Figure 3 approximately here)

Two coders (the second and third authors) labeled the same 96 graphs separately, and inter-rater reliability was calculated. Next, one of the two coders (the second author) coded an additional 583 graphs, giving an average of 8.1 graphs coded per student (not counting graphs used for calculating inter-rater reliability). Again, each graph corresponds to one student's entire span of activity for a specific knowledge component.

From the 583 graphs, an average occurrence of each tag for each student was computed. One tag, the constant tag, was never observed in the 583 graphs for which analysis was conducted, and thus was not used in further statistical analysis. The average occurrence of each tag was then correlated to each student's performance on each of the four post-tests of learning (straightforward problem-solving, transfer, preparation for future learning, and transfer). As this represents a substantial number of statistical analyses (6*4 = 24), we controlled for multiple comparisons. In specific, the analyses in this study utilize Storey et al.'s (2004) variant of the false discovery rate (FDR; Benjamini & Hochberg, 1995) method for hypothesis testing. This method produces a substitute for p-values, termed q-values, driven by controlling the proportion of false positives obtained via a set of tests. Whereas a p-value expresses that 5% of all tests may include false positives, a q-value indicates that 5% of significant tests may include false positives. As such, the FDR method does not guarantee each test's significance, but guarantees a low overall proportion of false positives, preventing the substantial over-conservatism found in methods such as the Bonferroni correction (cf. Perneger, 1998). The FDR calculations in the results section were made using the QVALUE software package (Storey, Taylor, & Siegmund, 2004) within the R statistical software environment (R Development Core Team, 2011).

Results

The coders' inter-rater reliability, across tags, was determined by computing the inter-rater reliability for each tag (e.g., whether that tag was included or not included), using Cohen's (1960) kappa. Kappa was then averaged across tags, to find average inter-rater agreement. Kappa is typically used for inter-rater reliability, because it controls for the possibility that agreement can occur by chance. A kappa of 0 indicates that coder agreement is equal to what would be expected from the base rate of each code, and a kappa of 1 indicates perfect agreement. Within this data set, the average kappa across constructs was 0.86, indicating an acceptable level of agreement between coders; Landis and Koch's (1977) guidelines for agreement suggest that agreement between 0.81 and 1.00 is "almost perfect."

(Place Table 3 approximately here)

Of the seven visual forms of the moment-by-moment learning curve, the most prevalent was immediate drop at 58.3%, followed by single spike at 24.8% and separated multi-spike at 14.5%. Other visual forms had lower prevalence, with close multi-spike at 9.6%, immediate peak at 8.2%, and plateau at 2.5%. The constant curve form never occurred in these graph replays. The prevalence of each visual form and their standard deviations are given in Table 3. We can infer from these results that the most frequent pattern of performance in this tutor was high initial performance (the immediate drop curve form), occurring a little more than half the time. Among the remaining students, about two thirds had a single spike, indicating a specific point where considerable learning occurred, or an immediate peak, indicating considerable learning early in the process. Only a minority had multiple spikes or a plateau.

Across the tests, most of the visual forms were statistically significantly associated with at least one of the tests of learning. The full set of results is presented in Table 4. Two visual forms were negatively significantly associated with one or more of the learning measures: close multi-spike and plateau. The plateau curve form, corresponding to considerable learning occurring over consecutive steps, was statistically significantly negatively associated with the problem-solving post-test, $r = -0.38$, $F(1,71) = 11.79$, $q < 0.01$; the transfer test, $r = -0.28$, $F(1,71) = 5.85$, $q = 0.04$; the PFL test, $r = -0.27$, $F(1,71) = 5.66$, $q = 0.04$; and the problem-solving retention test, $r = -0.52$, $F(1,71) = 25.65$, $q < 0.01$. The close multi-spike curve form, corresponding to close multiple periods of considerable learning, was marginally statistically significantly associated with the problem-solving post-test, $r = -0.25$, $F(1,71) = 4.61$, $q = 0.056$. One possible explanation for these findings is that these two visual forms – both indicating on multiple close events of learning – characterize students who started with initial low knowledge; recall that values on the moment-by-moment learning curves are proportional rather than absolute, which means that a student with low entry-level knowledge has more chance to gain knowledge during using the tutor than a student with high entry-level knowledge.

To investigate this explanation further, we can test the correlations between the frequency of the plateau and close multi-spike forms and students’ pretest scores. In doing so, we use Benjamini and Hochberg’s original procedure for using False Discovery Rate controls (Benjamini & Hochberg, 1995), as the QVALUE software package does not work for this small a number of tests. According to this procedure, only two relationships between a curve form and pre-test scores are statistically significant: the plateau form, r = -0.25, p < 0.05, and the close multi-spike form, r = -0.31, p < 0.01. Hence, students who had curve forms of these types likely had lower post-test score due to having lower initial knowledge.

Two visual forms were positively significantly associated with one or more of the post-test learning measures: immediate peak and immediate drop. The immediate drop curve form was statistically significantly associated with the problem-solving post-test, r = 0.32, F(1,71) = 7.93, q = 0.02; and the PFL test, r = 0.29, F(1,71) = 6.29, q = 0.04. The immediate peak curve form was statistically significantly positively associated with the problem-solving retention test, r = 0.35, F(1,71) = 9.73, q = 0.01. It was also marginally statistically significantly associated with the transfer test, r = 0.21, F(1,71) = 3.40, q = 0.098. These two forms correspond to either the student initially knowing the skill, or learning it very quickly and completely. As such, this finding might suggest relationships between quick learning and high achievements.

Two visual forms were not significantly associated, in either direction, with any of the four learning measures: single spike and separated multi-spike.

(Place Table 4 approximately here)

Discussion and Conclusions

Within this paper, we propose a new method for analyzing student learning over time, conducting visual analysis of the functional form of the moment-by-moment learning curve. The moment-by-moment learning curve is composed of measurements made by the moment-by-moment learning model (Baker, Goldstein, & Heffernan, 2010, 2011), an approach that combines Bayesian analysis and educational data mining to assess the probability that a student acquires a skill at a specific learning opportunity. The model is computed, and graphs are created which display a specific student's learning at each learning opportunity, for a specific skill. These graphs are then visually analyzed by human coders, in order to assess whether the graph possesses a set of specific, pre-defined visual forms. This application of a model developed through data mining to analyze a new research question is an example of the approach to using educational data mining which Baker and Yacef (2009) describe as "discovery with models." The original concept of the moment-by-moment learning curve was introduced in (Baker, Goldstein, & Heffernan, 2010, 2011), but in that work the curve was inspected solely in a qualitative fashion, except for a single mathematical measure (spikiness). The current study builds upon that previous work by discussing several characteristic visual forms exist for this curve, interpreting those forms' meaning, and analyzing the relationship between these forms and measures of robust learning, discovering that specific visual forms are significantly correlated with robust learning. We find that one visual form is particularly common, the "immediate drop" form, where a student shows high initial learning which reduces quickly. This visual form was present in over half of all graphs. The immediate drop visual form likely represents a student who already knows the relevant skill and simply must transfer the skill into the learning system; in some cases it may also represent a student immediately mastering a very easy skill. However, differences in the prevalence of the visual form between students can only be explained by the first explanation, that the student already knew the skill, as all students encountered the same skills within this data set. It is also worth noting that this visual form does not represent all cases where a student knows a skill in advance, as some cases along these lines would presumably show no learning at all (e.g., cases where the student's performance is rapid and flawless from the very beginning).

Among the visual forms, two were significantly positively associated with one or more of the measures of learning. The common immediate drop form was significantly positively associated with the problem-solving post-test. This relationship is unsurprising, as students who know key skills before using a tutor can be expected to perform better on a post-test of those skills. More surprisingly, however, the immediate drop form was also associated with better performance on the test of preparation for future learning. This finding may suggest that over-practice can lead students to not only develop greater speed of performance (Newell & Rosenbloom, 1981) and lower probability of forgetting (Pavlik & Anderson, 2005), but also the deeper conceptual knowledge required to prepare the student for future learning. However, given the links between over-practice and lower forgetting, it is somewhat surprising that the immediate drop visual form was not significantly associated with better performance on the retention test. The second visual form that was significantly positively associated with measures of learning was the immediate peak visual form. This visual form is similar to immediate drop, but has an initial lower learning followed by higher learning and then a drop. Due to the low initial learning, this visual form is less likely to represent knowledge being transferred in, though it may still represent a delayed realization of which knowledge is relevant to the current situation. This visual form is more likely to represent the student quickly learning the relevant skill. In a strong contrast, the plateau visual form was significantly negatively associated with all four measures of learning. This visual form represents students who have steady learning (e.g., steady improvement in performance) during only part of the learning activity.

Interestingly, visual forms involving spikes in learning were not negatively associated with learning to the same degree as the other visual forms. The single spike and the separated multi-spike visual forms were not significantly associated with any of the four measures of learning, and the close multi-spike form (the spike-based visual form most similar to the plateau visual form) was only marginally negatively associated with one of the four measures of learning. This finding suggests that, at least within this system, students who acquired learning more rapidly – in “eureka” moments punctuated by longer periods of stable performance – generally had better learning than students who manifested more gradual improvement. This finding cannot be completely explained at the current time, but has some interesting potential implications for research on learning. In specific, it argues that there are specific features of “eureka” moments of learning that lead to qualitatively different learning than the learning acquired more gradually. Studying what the cognitive differences are between these “eureka” moments may shed considerable light on important aspects of student learning. Past research on “eureka” moments on learning has typically focused on laboratory experiments using highly difficult problems thought to require a single insight for success. However, it has been argued – and we agree – that insight cannot be fully understood in these contexts (Bowden et al., 2005). Microgenetic analyses have provided evidence for factors that trigger qualitative shifts in performance, in specific situations (e.g., Siegler & Jenkins, 1989). Another possible interpretation of spikes – at least for multiple spikes – is that they represent the progressive deepening of a skill, where shallow understanding is replaced by deep understanding (cf. Cen et al., 2006). This may be another potential explanation for the better outcomes associated with spikes than with more gradual learning. Further analysis using the moment-by-moment learning model may be able to shed additional light on these possibilities, and support the development of greater understanding of the factors that elicit “eureka” moments, and the factors that lead to robust learning.

The model used in this paper to study changes in performance in a fine-grained fashion – that is, the moment-by-moment learning curve – has the potential to be used in other studies. In fact, it can be constructed for any learning system to which a Bayesian Knowledge Tracing (BKT; Corbett & Anderson, 1995) model can be applied. Generally, a BKT model can be applied for any computer-based learning environment in which the student is repeatedly interacting with the learning material in a way that the student's responses can be coded as "right"/"appropriate" or "wrong"/"inappropriate", with reference to a specific skill. Previously, BKT models have been used in various systems and for a range of domains, including mathematics (Koedinger & Corbett, 2006), scientific inquiry (Sao Pedro et al., in press), computer programming (Corbett & Anderson, 1995), reading (Mostow & Aist, 2001), and physics (Gertner & VanLehn, 2000). As the model and approach proposed here is generalizable to other learning systems, one direction for continuing this research would be to see whether the results found replicate in other systems, domains, and populations. Features of the learning system, as well as characteristics of the domain might play a role in students' learning, and the approach proposed here may eventually become a useful part of the methodological toolbox for studying such questions empirically.

Another future research direction would be to increase the degree to which the analyses presented here can be automatized. In their current form, human labels were required to assess the relevant visual forms. Given the human labels of the visual forms, it may now be possible to develop automated methods for inferring which visual form is present, using data mining. This step will speed future analyses of learning over time using the moment-by-moment learning model. In turn, it may even be possible to apply these models at run time, to facilitate assessment of the probability that a student will eventually acquire robust learning (cf. Baker, Gowda, &

Corbett, 2011). While models predicting robust learning already exist, the strong predictive relationships found in this paper argue that these models can be made significantly better through the inclusion of assessments of students' probable learning over time.

Acknowledgments

The authors would like to thank Ken Koedinger for helpful comments and suggestions, Albert Corbett for sharing the data set used in this paper and for helpful comments and suggestions, and Rob Martin for initial help in conceptualizing the visual forms of the moment-by-moment learning curve. This research was supported via grant "Empirical Research: Emerging Research: Robust and Efficient Learning: Modeling and Remediating Students' Domain Knowledge", National Science Foundation award number #DRL-0910188, and by the Pittsburgh Science of Learning Center, National Science Foundation award number #SBE-0836012.

References

Anderson, R.B., & Tweney, R.D. (1997). Artifactual power curves in forgetting. *Memory & Cognition*, 25(5), 724-730.

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

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

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. (2010). 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*, 52-63.

Baker, R.S.J.d., Corbett, A.T., Roll, I., & Koedinger, K.R. (2008). Developing a Generalizable Detector of When Students Game the System. *User Modeling and User-Adapted Interaction*, 18 (3), 287-314.

Baker, R.S.J.d., Corbett, A.T., & Wagner, A.Z. (2006). Human Classification of Low-Fidelity Replays of Student Actions. *Proceedings of the Educational Data Mining Workshop at the 8th International Conference on Intelligent Tutoring Systems*, 29-36.

Baker, R.S.J.d., Goldstein, A.B., & Heffernan, N.T. (2010). Detecting the Moment of Learning. *Proceedings of the 10th Annual Conference on Intelligent Tutoring Systems*, 25-34.

Baker, R.S.J.d., Goldstein, A.B., & Heffernan, N.T. (2011). Detecting learning moment-by-moment. *International Journal of Artificial Intelligence in Education, 21*(1-2), 5-25.

Baker, R.S.J.d., Gowda, S.M., & Corbett, A.T. (2011). Automatically Detecting a Student's Preparation for Future Learning: Help Use is Key. *Proceedings of the 4th International Conference on Educational Data Mining,* 179-188.

Baker, R.S.J.d., Pardos, Z., Gowda, S., Nooraei, B., & Heffernan, N. (2011). Ensembling Predictions of Student Knowledge within Intelligent Tutoring Systems. *Proceedings of 19th International Conference on User Modeling, Adaptation, and Personalization*, 13-24.

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 (1), 3-17.

Beck, J.E. (2006). Using learning decomposition to analyze student fluency development. In: *Proceedings of the Workshop on Educational Data Mining at the 8th International Conference on Intelligent Tutoring Systems*, 21-28. Jhongli, Taiwan.

Beck, J. E., & Mostow, J. (2008). How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, 353-362.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289–300.

Bowden, E.M., Jung-Beeman, M., Fleck, J., & Kounios, J. (2005). New approaches to demystifying insight. *Trends in Cognitive Sciences*, 9(7), 322-328.

Bower, G.H. (1961). Application of a model to paired-associate learning. *Psychometrika*, *26*(3), 255-280.

Bower, G.H. & Trabasso, T. (1963). Reversals prior to solution in concept identification. *Journal or Experimental Psychology*, *66*(4), 409-418.

Bransford, J. D., & Schwartz, D. (1999). Rethinking transfer: A simple proposal with multiple implications. *Review of Research in Education*, 24, 61-100.

Cen, H., Koedinger, K.R., & Junker, B. (2006). Learning Factors Analysis: A general method for cognitive model evaluation and improvement. In M. Ikeda, K. Ashley, & T. Chan (Eds.),*Intelligent Tutoring Systems 8th International Conference* (pp. 164–175). Berlin: Springer.

Cen, H., Koedinger, K.R., & Junker, B. (2007). Is Over Practice Necessary? – Improving Learning Efficiency with the Cognitive Tutor using Educational Data Mining. In Lucken, R., Koedinger, K. R. and Greer, J. (Eds). *Proceedings of the 13th International Conference on Artificial Intelligence in Education*, pp. 511-518.

Cobb, P. (1999). Individual and Collective Mathematical Development: The Case of Statistical Data Analysis. *Mathematical Thinking and Learning*, 1(1), 5-43.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20* (1), 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.

Corbett, A.T., MacLaren, B., Kauffman, L., Wagner, A., & Jones, E. (2010). A Cognitive Tutor for genetics problem solving: Learning gains and student modeling. *Journal of Educational Computing Research*, 42, 219-239.

Corbett, A., MacLaren, B., Wagner, A., Kauffman, L., Mitchell, A., Baker, R.S.J.d., & Gowda, S.M. (2011). Preparing Students for Effective Explaining of Worked Examples in the Genetics Tutor. *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*, 1476-1481.

D'Mello, S., & Graesser, A. (2011). The half-life of cognitive-affective states during complex learning. *Cognition & Emotion*, 25(7), 1299-1308.

Estes, W.K. (1950). Toward a statistical theory of learning. *Psychological Review*, *57*(2), 94-107.

Fong, G.T., & Nisbett, R.E. (1991). Immediate and delayed transfer of training effects in statistical reasoning. *Journal of Experimental Psychology: General, 120,* 34–45.

Gertner, A.S., & VanLehn, K. (2000). Andes: A coached problem solving environment for Physics. In *Proceedings of the 5th International Conference on Intelligent Tutoring Systems*, 133-142.

Hall, M.A. (2000). Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning. *Proceedings of the 17th International Conference on Machine Learning*, 359-366.

Heathcote, A., Brown, S., & Mewhort, D,J.K. (2000). The Power Law Repealed: The Case for an Exponential Law of Practice. *Psychonomic Bulletin and Review*, 7, 185-207.

Henderson, K. (1999) *On line and on paper: Visual representations, visual culture, and computer graphics in design engineering*. Cambridge, MA: MIT Press.

Hershkovitz, A., & Nachmias, R. (2009). Learning about online learning processes and students' motivation through web usage mining. *Interdisciplinary Journal of E-Learning and Learning Objects*, 5, 197-214.

Koedinger, K. R., & Corbett, A. T. (2006). Cognitive Tutors: Technology bringing learning science to the classroom. In K. Sawyer (Ed.) *The Cambridge Handbook of the Learning Sciences*, 61-78. Cambridge University Press.

Koedinger, K.R., Corbett, A.T., & Perfetti, C. (in press). The Knowledge-Learning-Instruction (KLI) Framework: Toward Bridging the Science-Practice Chasm to Enhance Robust Student Learning. To appear in *Cognitive Science.*

Kuhn, D., Goh, W., Iordanou, K., & Shaenfield, D. (2008). Arguing on the computer: A microgenetic study of developing argument skills in a computer-supported environment. *Child Development, 79*(5), 1310-1328.

Kuhn, D., Schauble, L., & Garcia-Mila, M. (1992). Cross-domain development of scientific reasoning. *Cognition and Instruction*, 9, 285-327.

Landis J.R., & Koch G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.

Lehrer, R., Schauble, L., Strom, D., & Pligge, M. (2001). Similarity of form and substance: Modeling material kind. In D. Klahr & S. Carver (Eds.), *Cognition and instruction: 25 years of progress*. (pp. 39-74). Mahwah, NJ: Lawrence Erlbaum Associates.

Lindstrom, P., & Gulz, A. (2008). Catching Eureka on the Fly. In: *Proceedings of the AAAI 2008 Spring Symposium*.

Martin, B., Mitrovic, A., Koedinger, K.R., & Marthan, S. (2011). Evaluating and improving adaptive educational systems with learning curves. *User Modeling and User-Adapted Interaction*, 21(3), 249-283.

Mathan, S.A., & Koedinger, K.R. (2005). Fostering the Intelligent Novice: Learning From Errors With Metacognitive Tutoring. *Educational Psychologist*, 40(4), 257-265.

Mazur, J. E., & Hastie, R. (1978). Learning as accumulation: A reexamination of the learning curve. *Psychological Bulletin, 85*(6), 1256-1274.

Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., & Euler, T. (2006). YALE: Rapid Prototyping for Complex Data Mining Tasks. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD 2006), 935-940.

Mostow, J., & Aist, G. (2001). Evaluating tutors that listen: An overview of Project LISTEN. In K.D. Forbus (Ed.) & Feltovich, P.J. (Eds.), *Smart Machines in Education: The coming revolution in educational technology (pp. 169-234)*. Cambridge, MA: The MIT Press.

Newell, A., & Rosenbloom, P.S. (1981). Mechanisms of Skill Acquisition and the Law of Practice. In J.R. Anderson (Ed.) *Cognitive Skills and their Acquisition*, 1-55. Hillsdale, NJ: Lawrence Erlbaum Associates.

Pavlik, P.I., & Anderson, J.R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, 29, 559–586.

Perneger, T.V. (1998). What's wrong with Bonferroni adjustments. *British Medical Journal*, *316*, 1236-1238.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Romero, C., & Ventura, S. (2007). Educational Data Mining: A Survey from 1995 to 2005. *Expert Systems with Applications*, 33 (1), 135-146.

Sao Pedro, M. A., Baker, R. S. J. d., Montalvo, O., Nakama, A., & Gobert, J. D. (2010). Using text replay tagging to produce detectors of systematic experimentation behavior patterns. *Proceedings of the 3rd International Conference on Educational Data Mining,* 181-190.

Sao Pedro, M.A., Baker, R.S.J.d., Gobert, J., Montalvo, O., & Nakama, A. (in press). Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. To appear in *User Modeling and User-Adapted Interaction*.

Schmidt, R.A., & Bjork, R.A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3,* 207-217.

Siegler, R.S., & Crowley, K. (1991). The microgenetic method: A direct means for studying cognitive development. *American Psychologist*, 46(6), 606-620.

Siegler, R.S., & Jenkins, E. (1989). *How children discover new strategies*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Singley, M.K., & Anderson, J.R. (1989). *The Transfer of Cognitive Skill*. Cambridge, MA: Harvard University Press.

Storey, J.D., Taylor, J.E., & Siegmund, D. (2004). Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B*, 66(1), 187-205.

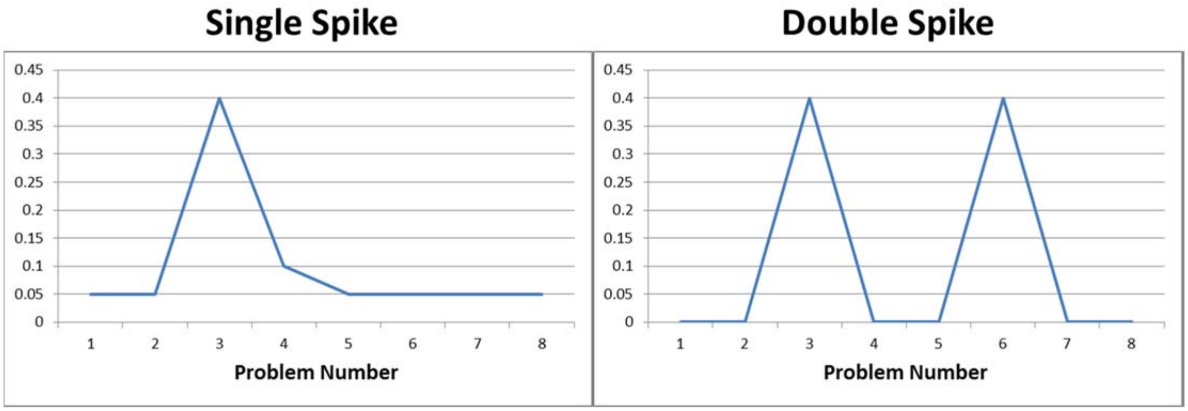
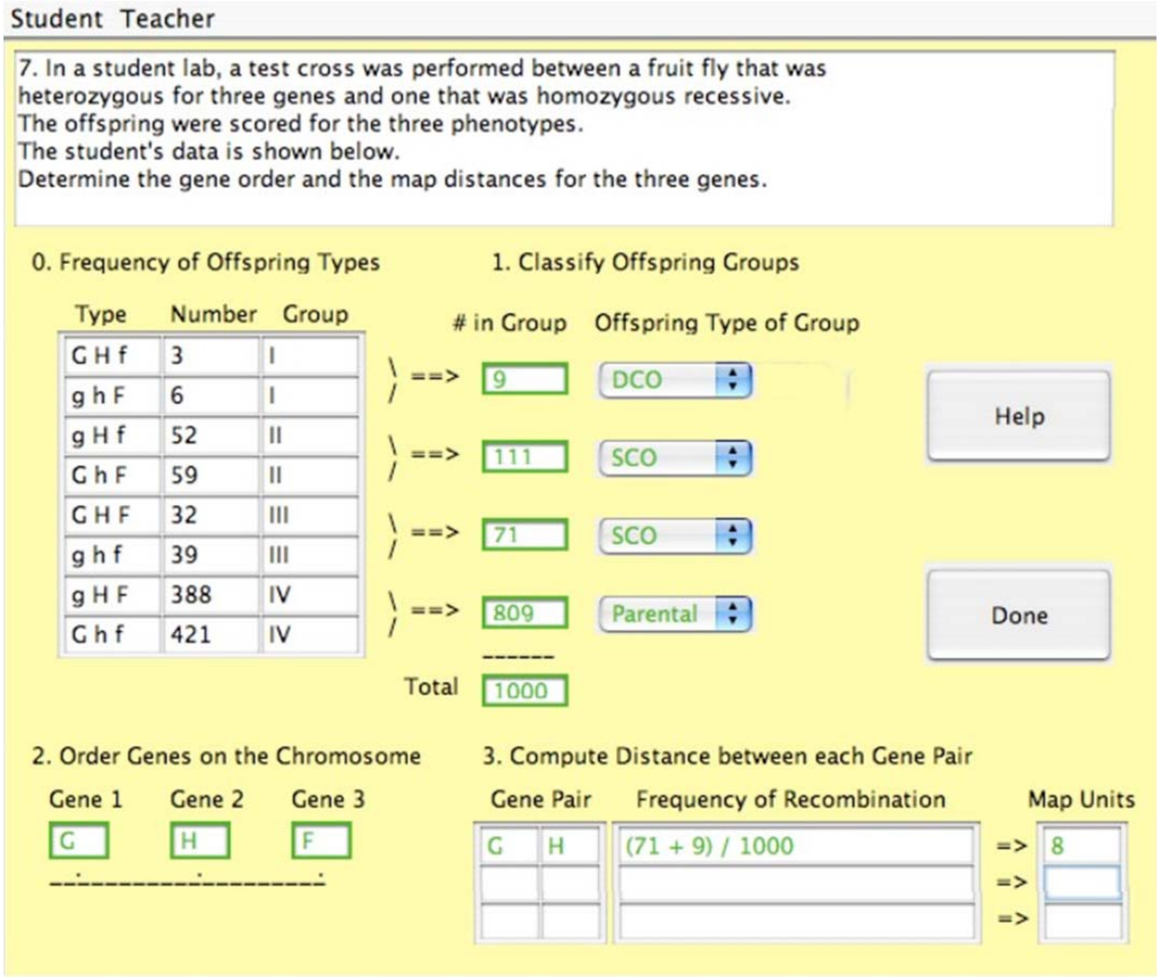


Figure 1 – Two graphs with equal spikiness exhibiting two different patterns of learning.



7. In a student lab, a test cross was performed between a fruit fly that was heterozygous for three genes and one that was homozygous recessive. The offspring were scored for the three phenotypes. The student's data is shown below. Determine the gene order and the map distances for the three genes.

0. Frequency of Offspring Types

Type	Number	Group
G H f	3	I
g h F	6	I
g H f	52	II
G h F	59	II
G H F	32	III
g h f	39	III
g H F	388	IV
G h f	421	IV

1. Classify Offspring Groups

# in Group	Offspring Type of Group
9	DCO
111	SCO
71	SCO
809	Parental

Total 1000	

2. Order Genes on the Chromosome

Gene 1	Gene 2	Gene 3
G	H	F

3. Compute Distance between each Gene Pair

Gene Pair		Frequency of Recombination		Map Units
G	H	(71 + 9) / 1000	=>	8
			=>	
			=>	

Figure 2 – A screenshot from the Three-Factor Cross lesson of the Genetics Cognitive Tutor

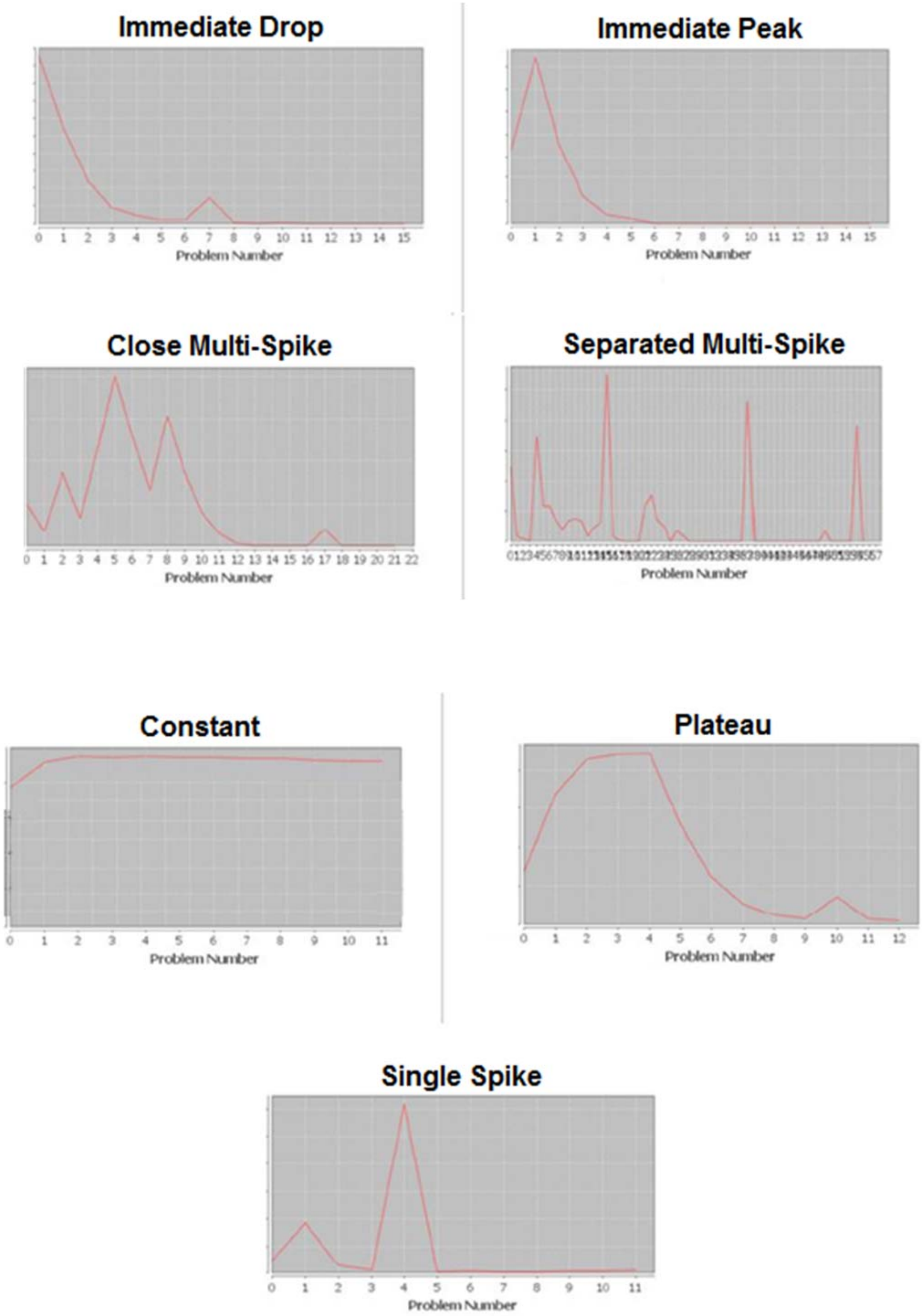


Figure 3 – Examples of visual forms of the moment-by-moment learning curve used in analysis.

Table 1 – Correlations between the post-test measures of student learning.

	Transfer	PFL	Retention
Problem-solving	0.59	0.41	0.45
Transfer		0.52	0.48
PFL			0.33

Table 2 – The best-fitting linear regression model of P(***J***). Values of P(***J***) can be computed by multiplying each feature by its coefficient (on the right) and summing all values together.

Feature	**P(J) =**
The action is assessed by the learning system as correct	-0.0069
The action is assessed by the learning system as incorrect	+0.0069
Action is a help request	- 0.0270
Action is assessed as a misconception by the learning system	+0.0511
Action involves typing a string	-0.0412
Action involves typing a number	-0.0412
Time taken (SD faster (-) or slower (+) than average across all students)	-1.3688
Time taken in last 3 actions (calculated in SD off average across students)	-0.6220
Time taken in last 5 actions (calculated in SD off average across students)	+0.7557
The number of times the student asked for help at this skill, including previous problems	-0.0070
The number of times the student made errors at this skill, including previous problems	-0.0002
Number of last 3 actions which involved same interface element	-0.0165
Number of last 5 actions which involved same interface element	+0.0051
Number of opportunities student has already had to use current skill	-0.0001
The probability the student knew the skill, after the current action (Ln)	-0.0195
The probability the student knew the skill, before the current action (Ln-1)	-0.0960
Constant	+0.1632

Table 3 – Average prevalence of each visual form of the moment-by-moment learning curve.

	Average Prevalence	Standard Deviation
Single Spike	24.8%	14.9%
Close Multi-Spike	9.6%	9.7%
Separated Multi-Spike	14.5%	11.9%
Plateau	2.5%	6.2%
Constant	0.0%	0.0%
Immediate Peak	8.2%	10.1%
Immediate Drop	58.3%	23.3%

Table 4 – The correlation between a student’s proportion of a specific visual form of the moment-by-moment learning curve across skills, and their performance on the four learning tests. Statistically significant findings (controlling for false discovery rate) are highlighted in dark gray; marginally significant findings are highlighted in light gray.

Curve form	**Test**	**r**	**F**	**p**	**q**
pct single spike	Post-test	0.075	0.400	0.529	0.374
	Transfer test	-0.036	0.095	0.759	0.446
	PFL test	-0.139	1.402	0.240	0.253
	Retention Test	-0.094	0.636	0.428	0.330
pct close multi-spike	*Post-test*	*-0.247*	*4.610*	*0.035*	*0.056*
	Transfer test	-0.094	0.634	0.429	0.330
	PFL test	-0.035	0.085	0.771	0.446
	Retention Test	0.045	0.142	0.708	0.446
pct separated multi-spike	Post-test	-0.134	1.301	0.258	0.253
	Transfer test	0.011	0.008	0.927	0.492
	PFL test	-0.113	0.916	0.342	0.311
	Retention Test	0.063	0.285	0.595	0.399
pct plateau	**Post-test**	**-0.377**	**11.786**	**0.001**	**0.006**
	Transfer test	**-0.276**	**5.847**	**0.018**	**0.036**
	PFL test	**-0.272**	**5.663**	**0.020**	**0.036**
	Retention Test	**-0.515**	**25.647**	**0.000**	**0.000**
pct immediate peak	Post-test	0.092	0.601	0.441	0.330
	Transfer test	*0.214*	*3.399*	*0.069*	*0.098*
	PFL test	0.017	0.021	0.886	0.490
	Retention Test	**0.347**	**9.725**	**0.003**	**0.011**
pct immediate drop	**Post-test**	**0.317**	**7.930**	**0.006**	**0.020**
	Transfer test	0.167	2.035	0.158	0.183
	PFL test	**0.285**	**6.286**	**0.014**	**0.036**
	Retention Test	0.206	3.152	0.080	0.102