Applying Three Models of Learning to Individual Student Log Data

Brett van de Sande
Arizona State University
PO Box 878809
Tempe, AZ 85287
bvds@asu.edu

ABSTRACT

Normally, when considering a model of learning, one compares the model to some measure of learning that has been aggregated over students. What happens if one is interested in individual differences? For instance, different students may have received different help, or may have behaved differently. In that case, one is interested in comparing the model to the individual learner. In this study, we investigate three models of learning and compare them to student log data with the goal of seeing which model best describes individual student learning of a particular skill. The log data is from students who used the Andes intelligent tutor system for an entire semester of introductory physics. We discover that, in this context, the "best fitting model" is not necessarily the "correct model" in the usual sense.

Keywords

data mining, models of student learning

1. INTRODUCTION

Most Knowledge Component (KC) [15] based models of learning are constructed in a similar manner, following Corbett and Anderson [8]. First, some measure of learning is selected (*e.g.* correct/incorrect on first try) for the j-th opportunity for that student to apply a given KC. This measure of learning is then aggregated over students (*e.g.* fraction of students correct) as a function of j. Finally, aggregated measure is then compared to some model (*e.g.* Bayesian Knowledge Tracing) with model parameters chosen to optimize the model's fit to the data. In principle, given sufficient student log data, one could uniquely determine which of several competing models best matches the data.

One drawback with this approach is that it does not take into account individual learner differences or the actual behaviors of students or tutors as they are learning. Thus, a number of authors have extended their models to include individual student proficiency and actual help received by the student. For instance, in the Cordillera natural language tutoring system for physics [16], the student may have been asked what the next step was or were told what the next stop was; this was used as input for an associated model. An overview of these models can be found in [7].

If one is primarily interested in the effectiveness of help given to an individual student or the effectiveness (for learning) of a particular strategy or behavior of a student, then it may make sense to fit a model of learning to the log data of each student individually. Given sufficient student log data, can we still talk about a particular model fitting the student log data well? That is the central question of this paper. To start our investigation, we will compare three different models of learning using data from students taking introductory physics and examine whether there is empirical support for using one model over the others. In fact, using Akaike Information Criteria (AIC), we obtain results that seem to favor two models over the third, but note that fitting the models to individual students can make the determination ambiguous.

1.1 Correct/Incorrect steps

Our stated goal is to determine student learning for an individual student as they progress through a course. What observable quantities should be used to determine student mastery? One possible observable is "correct/incorrect steps," whether the student correctly applies a given skill at a particular problem-solving step without any preceding errors or hints. There are other observables that may give us clues on mastery: for instance, how much time a student takes to complete a step that involves a given skill. However, other such observables typically need some additional theoretical interpretation. *Exempli gratia*, What is the relation between time taken and mastery? Baker, Goldstein, and Heffernan [3] develop a model of learning based on a Hidden Markov model approach. They start with a set of 25 additional observables (including "time to complete a step") and construct their model and use correct/incorrect steps to calibrate the additional observables and determine which are significant. Naturally, it is desirable to eventually include various other observables in any determination of student learning. However, in the present investigation, we will focus on correct/incorrect steps.

Next, we need to define precisely what we mean by a step. A student attempts some number of *steps* when solving a problem using an intelligent tutor system (ITS). Usually, a step

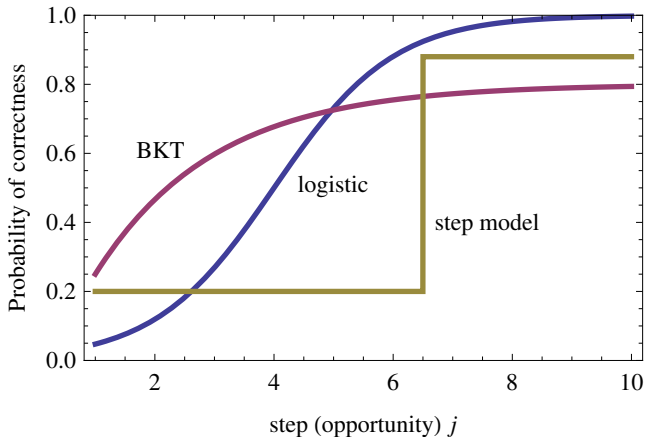

Figure 1: Functional form of the three models of student learning.

is associated with creating/modifying a single user interface object (writing an equation, drawing a vector, defining a quantity, *et cetera*) and is a distinct part of the problem solution (that is, help-giving dialogs are not considered to be steps). A student may attempt a particular problem-solving step, delete the object, and later attempt that solution step again. A step is an *opportunity* to learn a given Knowledge Component (KC) [15] if the student must apply that skill to complete the step.

Andes is a model-tracing tutor [2], which means that the ITS contains a number of "model solutions" to each problem and each step of the model solution has one or more KCs assigned to it. As a student solves a problem, Andes tries to match each student attempt at a step to a corresponding model solution step and, if that match is successful, assigns the corresponding KCs to that step attempt. For some common errors, Andes has a number of error detectors that infer what solution step the student was attempting to work on. In that case, KCs can assigned to that attempt. However, there are many errors where the associated KCs cannot be determined. In the log analysis, if a step attempt does not have any KCs assigned to it, we use the following heuristic to determine the associated KCs: First, we look at any subsequent attempts associated with the same user interface element and see if they have any KCs associated with them. If that fails, then we look for the next attempt having the same *type* of user interface element (equation, vector, *et cetera*) that has some KCs associated with it.

For each KC and student, we select all attempted steps that involve application of that KC and mark each step as "correct" if the student completes that step correctly without any preceding errors or requests for help; otherwise, we mark the step as "incorrect." If each incorrect/correct step is marked with a 0/1, then a single student's performance on a single KC can be expressed as a bit sequence, *exempli gratia* 00101011. We will label steps with $j \in \{1, \ldots, n\}$.

2. THREE MODELS OF LEARNING

Ultimately, we are interested in determining when a student has mastered a particular KC and, by inference, the effectiveness of any help given by the tutor. Thus, a useful model of learning should have the the following properties:

1. Be compatible with actual student behavior. That is, its functional form should fit well with student data. We will explore this question in Section 3.
2. Give the probability that learning has occurred at a given step.
3. Assuming learning has occurred at a given step, the model should give a prediction for the associated increase in performance and the rate of errors after learning.

We will consider three candidate models: the Bayesian Knowledge Tracing (BKT) model, the logistic function, and the "step model;" see Fig. 1.

The first model is the Bayesian Knowledge Tracing (BKT) model [8]. The hidden Markov model form of BKT is often fit to student performance data [4]. One can show that this model, in functional form, is an exponential function with three model parameters [13]:

$$P_{\mathrm{BKT}}(j) = 1 - P(S) - Ae^{-\beta j} \ . \tag{1}$$

One central assumption of BKT is that, given that learning has not already occurred, mastery is *equally probable* on each step. This assumption of equal probability does not match well with our goal of determining empirically the steps where learning has actually occurred for an individual student, criterion 2. On the other hand, this model does provide the final error rate $P(S)$ (the initial error rate is ambiguous), so criterion 3 is partially satisfied.

A number of models of learning based on logistic regression have been studied [6, 10, 7]. These models involve fitting data for multiple students and multiple KCs and may involve other observables such as the number of prior successes/failures a student has had for a given skill. However, in this investigation, we are interested in fitting to the correct/incorrect bit sequence for a single student and a single KC and a logistic regression model takes on a relatively simple form

$$\log\left(\frac{P_{\mathrm{logistic}}(j)}{1 - P_{\mathrm{logistic}}(j)}\right) = b(j - L) \tag{2}$$

which can be written as:

$$P_{\mathrm{logistic}}(j) = \frac{1}{1 + \exp\left(-b(j - L)\right)} \ . \tag{3}$$

It is natural to associate L with the moment of learning. However, the finite slope of $P_{\mathrm{logistic}}(j)$ means that learning may occur in a range of roughly $1/b$ steps before and after L. For $P_{\mathrm{logistic}}(j)$, the gain in performance is always 1 and the final error rate is always 0. Thus, although this model makes a prediction for when the skill is learned, criterion 2, it does not predict a gain in performance, criterion 3.

The third model is the "step model" which assumes that learning occurs all at once; this corresponds to the "eureka learning" discussed by [3]. It is defined as:

$$P_{\mathrm{step}}(j) = \begin{cases} g, & j < L \\ 1 - s, & j \geq L \end{cases} \tag{4}$$

where L is the step where the student first shows mastery of the KC, g is the "guess rate," the probability that the student gets a step correct by accident, and s is the "slip rate," the chance that the student makes an error after learning the skill. These are analogous to the guess and slip parameters of BKT [8]. The associated gain in performance is $1-g-s$ and the error rate after learning is simply s in this model. Thus, this model satisfies criteria 2 and 3.

3. MODEL SELECTION USING AIC

The BKT and logistic function models are widely used and we have introduced the step model $P_{\text{step}}(j)$ as an alternative. How well do these models match actual student behavior? Since we will use the step model in subsequent work, it would be reassuring to know whether it describes the student data as well (or better than) the other two models. We will use the Akaike Information Criterion (AIC) for this purpose [1, 5]. AIC is defined as

$$\text{AIC} = -2\log(\mathcal{L}) + 2K \tag{5}$$

where $\mathcal{L}$ is the maximized value of the likelihood function and K is the number of parameters in the model. AIC is an estimate of the expected relative "distance" between a given model and the true model (assumed to be complicated) that actually generated the observed data. It is valid in limit of many data points, $n \to \infty$, with leading corrections of order $1/n$.

A related method for choosing between models is the Bayesian Information Criterion (BIC) introduced by Schwarz [11]. BIC is defined as

$$\text{AIC} = -2\log(\mathcal{L}) + K\log(n) \tag{6}$$

where n is the number of data points. Burnham & Anderson [5, Sections 6.3 & 6.4] explain that BIC is more appropriate in cases where the "true" model that actually created the data is relatively simple (few parameters). If the true model is contained in the set of models being considered, then BIC will correctly identify the true model in the $n \to \infty$ limit. For BIC to have this property, the true model must stay fixed as n increases. The authors argue that, while BIC may be appropriate in some of the physical sciences and engineering, in the biological and social sciences, medicine, and other "noisy" sciences, the assumptions that underlie BIC are generally not met. In particular, as the sample size increases, it is typical that the underlying "true" model also becomes more complicated. This is certainly true in educational datamining: datasets are generally increased by adding data from new schools, or different years and one generally expects noticeable variation of student behavior from school to school or from year to year. In such cases, one safely can say that the "true" model is complicated (because people are complicated) and becomes more complicated as a dataset is increased in size. Although most authors quote both AIC and BIC values, there is good reason to believe that AIC is generally more appropriate for educational datamining work.

3.1 Method

We examined log data from 12 students taking an intensive introductory physics course at St. Anselm College during summer 2011. The course covered the same content as a normal two-semester introductory course. Log data was recorded as students solved homework problems while using the Andes intelligent tutor homework system [17]. 231 hours of log data were recorded. Each student problem-solving step is assigned one or more KCs using the heuristic described in Section 1.1. The dataset contains a total of 2017 distinct student-KC sequences covering a total of 245 distinct KCs. We will refer to this dataset as student dataset $\mathcal{A}$. See Figure 2 for a histogram of the number of student-KC sequences having a given number of steps.

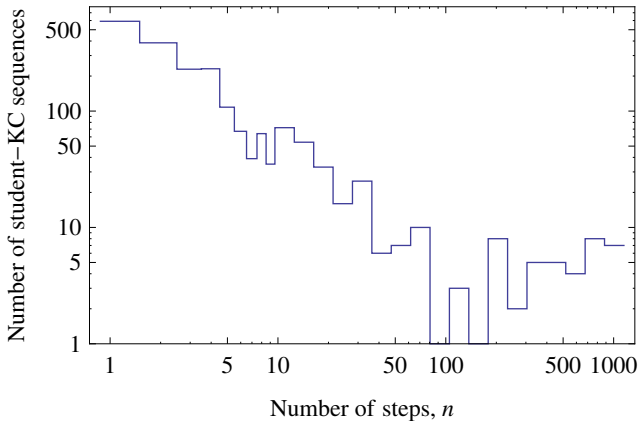

Figure 2: Histogram of number of distinct student-KC sequences in student dataset $\mathcal{A}$ having a given number of steps n.

Most KCs are associated with physics or relevant math skills while others are associated with Andes conventions or user-interface actions (such as, notation for defining a variable). The student-KC sequences with the largest number of steps are associated with user-interface related skills, since these skills are exercised throughout the entire course.

One of the most remarkable properties of the distribution in Fig. 2 is the large number of student-KC sequences containing just a few steps. The presence of many student-KC sequences with just one or two steps may indicate that the default cognitive model associated with this tutor system may be sub-optimal; to date, there has not been any attempt to improve on the cognitive model of Andes with, say, Learning Factors Analysis [6]. Another contributing factor is the way that introductory physics is taught in most institutions, with relatively little repetition of similar problems. This is quite different than, for instance, a typical middle school math curriculum where there are a large number of similar problems in a homework assignment.

3.2 Analysis

Since the goodness of fit criterion, AIC, is valid in the limit of many steps, we include in this analysis only student-KC sequences that contain 10 or more steps, reducing the number of student-KC sequences to 267, covering 38 distinct KCs. We determine the correctness of each step (Section 1.1), constructing a bit sequence, *exempli gratia* 001001101, for each student-KC sequence. This bit sequence is then fit to each of the three models, P_{step}, P_{logistic}, and P_{BKT} by maximiz-

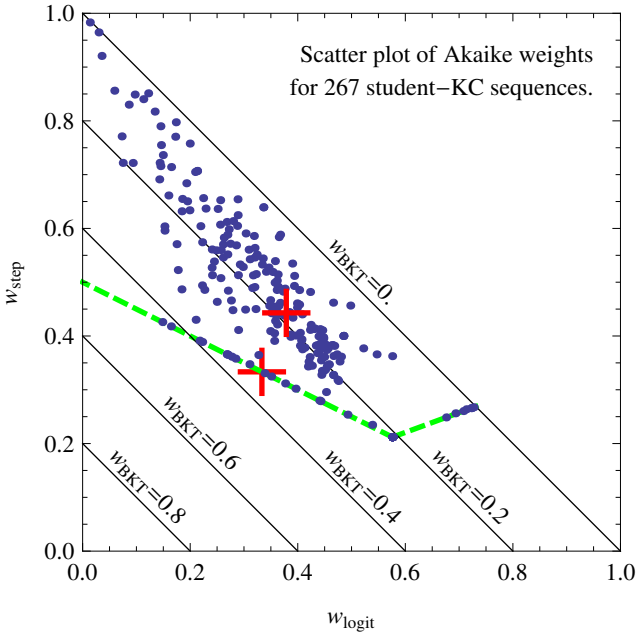

Figure 3: Scatter plot of Akaike weights for the three models, $P_{\rm step}$, $P_{\rm logistic}$, and $P_{\rm BKT}$, when fit to student-KC sequences from an introductory physics course. The point where all models are equal, $w_{\rm step} = w_{\rm logistic} = w_{\rm BKT} = 1/3$, is marked with the lower cross. The average of the weights is marked with the upper cross. The dashed line on the left represents points where $w_{\rm step} = w_{\rm BKT}$. Finally, the dashed line on the right marks data with bit sequences of the form $00\cdots011\cdots1$.

ing the associated log likelihood. For $P_{\rm logistic}$, and $P_{\rm BKT}$, the fits were calculated using the Differential Evolution algorithm [12] provided by *Mathematica*. For $P_{\rm step}$, the best fit, as a function of s and g, can be found analytically; one can then find the best fit, as a function of L, by conducting an exhaustive search. Next, we calculate the AIC score for each fit. Finally, we calculate the Akaike weights, $w_{\rm logistic}$, $w_{\rm step}$, and $w_{\rm BKT}$ for each student-KC sequence [5]. The weights are normalized so that

$$1 = w_{\rm logistic} + w_{\rm step} + w_{\rm BKT} \ . \tag{7}$$

The Akaike weight represents the relative probability that a particular model in a given set of models is closest to the model that has actually generated the data.

A scatter plot of the weights is shown in Fig. 3. If all three models described the data equally well, then we would expect points to be scattered evenly about the center point $w_{\rm logistic} = w_{\rm step} = w_{\rm BKT} = 1/3$. Instead, we see the step model (average weight 0.44) weakly favored over the logistic model (average weight 0.37) and strongly favored over BKT (average weight 0.18). Indeed, we find no data points where $w_{\rm step} < w_{\rm BKT}$, although there is a noticeable accumulation of points along the line $w_{\rm step} = w_{\rm BKT}$.

Note that data in the form of incorrect steps then correct steps, *exempli gratia* $00\cdots011\cdots1$, is fit perfectly by both the $P_{\rm step}$ and $P_{\rm logistic}$ models. In this case, since $P_{\rm logistic}$ has one fewer parameter than $P_{\rm step}$, it is favored by AIC by a constant factor and $w_{\rm step} = {\rm e}^{-1} w_{\rm logistic}$. This case is plotted as the increasing dashed line in Fig. 3.

Since the student-KC sequences contain an average of about $n = 16$ steps, it is surprising that we find that AIC so strongly discriminates between the models. Perhaps, though, this is due to some finite n correction: recall that AIC is only strictly valid in the $n \to \infty$ limit.

3.3 Random data

To further investigate the observed strong discrimination between the three models, we constructed an artificial dataset containing random bit sequences (each step has 50% probability of being "correct") of length $n \in \{10, 20, 30, 40, 50\}$, with 10,000 sequences for each n. This dataset corresponds to a model of the form

$$P_{\rm random}(j) = 1/2 \ . \tag{8}$$

We then repeated our analysis of the three models using this dataset and AIC as our selection criterion. Note that all three models, with a suitable choice of parameters, can be made equal to $P_{\rm random}$ itself.

As mentioned earlier, for data that is generated by a simple model (and $P_{\rm random}$ is about as simple as one can get) and the "true" model is included among the set of models, BIC is the more appropriate criterion for model selection [5, Sections 6.3 & 6.4]. However, for our results, the only difference between AIC and BIC is that BIC favors $P_{\rm logistic}$ more strongly over $P_{\rm step}$ and $P_{\rm BKT}$. Thus, using BIC would shift the weights so that $w_{\rm logistic}$ would increase somewhat over the other two weights. However, in order to maintain consistency with our experimental results, Fig. 3, we used AIC for the random data as well; see Fig. 4. This use of AIC versus BIC does not affect our conclusions.

For data generated by $P_{\rm random}$, one expects that all three models should perform equally well since all three can equal (with suitable choice of parameters) the known correct model $P_{\rm random}$. Thus, we would expect a scatter plot of the Akaike weights to center around $w_{\rm logistic} = w_{\rm step} = w_{\rm BKT} = 1/3$. Instead, we find that $P_{\rm step}$ is still highly favored over the other two; see Fig. 4. This bias seems to persist as we increase n.

Since we know that AIC (or BIC) is only strictly valid in the asymptotic limit $n \to \infty$, it is useful to see if the large differences persist as n is increased. If we average over the 10,000 weights and plot the average weight as a function of n, we find that the differences between the weights persist in the $n \to \infty$ limit; see Fig. 5. If we fit the average weights to a constant plus $1/n$; the fits are:

$$\langle w_{\rm step} \rangle = 0.58 - \frac{1.50}{n} \tag{9}$$

$$\langle w_{\rm logistic} \rangle = 0.24 + \frac{1.2}{n} \tag{10}$$

$$\langle w_{\rm BKT} \rangle = 0.17 + \frac{0.30}{n} \ . \tag{11}$$

This shows that AIC, in the asymptotic limit $n \to \infty$, still favors $P_{\rm step}$ over the other two models when used to evaluate

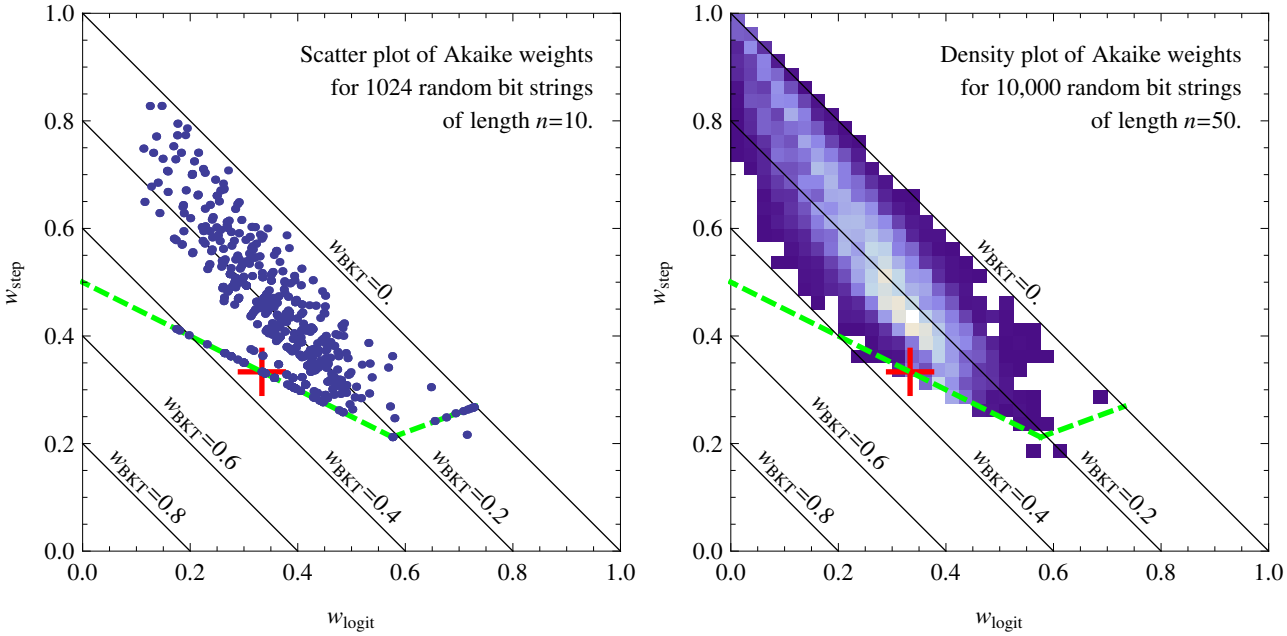

Figure 4: Akaike weights for the three models, P_{step}, P_{logistic}, and P_{BKT}, when fit to randomly generated data. The point where $w_{\text{step}} = w_{\text{logistic}} = w_{\text{BKT}} = 1/3$ is marked with a cross. For these datasets, each model should perform equally well, since, with an appropriate choice of parameters, they all can be made equal to the model that was used to generate the data.

randomly generated data.

If we repeat this analysis with BIC, we would still find that the weights converge to a constant value with $1/n$ leading errors. The only difference is that the logistic model has a larger weight than the other two. The differences between the weights of the three models still persist in the $n \to \infty$ limit.

3.4 Conclusions

In conclusion, we obtain some surprising results when we compare the three models, P_{step}, P_{logistic}, and P_{BKT}, using individual student data. We see that AIC weakly favors the step model over the logistic model in a fashion that one might expect. However, in an unexpected fashion, we see that both are strongly favored over the BKT model. We see that this effect persists for randomly generated data and is not due to an insufficient number of opportunities (finite n effect).

Moreover, for any bit sequence, P_{BKT} *never* fits the data better than P_{step}. Since both models have three parameters, this result holds for any maximum likelihood-based criterion, including both AIC and BIC. We don't have an analytic proof for this result, but the numerical evidence (see Fig. 5) is quite strong. In other words, even if one uses P_{BKT} (for some set of model parameters) to generate a bit sequence, one can always adjust the parameters in P_{step} so that it fits the bit sequence as well as, or better than, P_{BKT}.

What does this mean? Let us think more carefully about maximum likelihood. If one uses a model to generate a single bit sequence, we cannot determine the exact probability function (the probability as a function of j) that generated it. At best, one can only talk about the probability that given a function *may* have generated that sequence. On the other hand, if one uses a particular probability function to generate a collection of infinitely many sequences, then we know the exact probability for each step. Therefore, given the collection of many sequences, one can uniquely determine the probability function that generated that collection. If that function comes from a particular model $\mathcal{A}$ (for some choice of model parameters), then we can safely conclude that model $\mathcal{A}$ is the correct model.

In other words, when we fit individual student data to a model (fitting model parameters separately for each student), then we can make no statements about what model is "correct" in the sense that it may have generated the data. We can only talk about a model being a good fit in the sense that it is "close" to the data. On the other hand, if we aggregate data from many students and fit to a model (finding the best fit model parameters), then we can talk about a model being correct in the usual sense that it may have generated the data.

If we are interested in determining the effectiveness of help given or of a particular student behavior, we are more concerned about being "close" to the student data than finding the correct theory of learning, so the fact that the step model fits the data better than the logistic function and the BKT model is of practical value when analyzing student log data. However, one should not then conclude that the step function is a better model of student learning, in the usual sense. The better fit does not predict anything about the nature of learning.

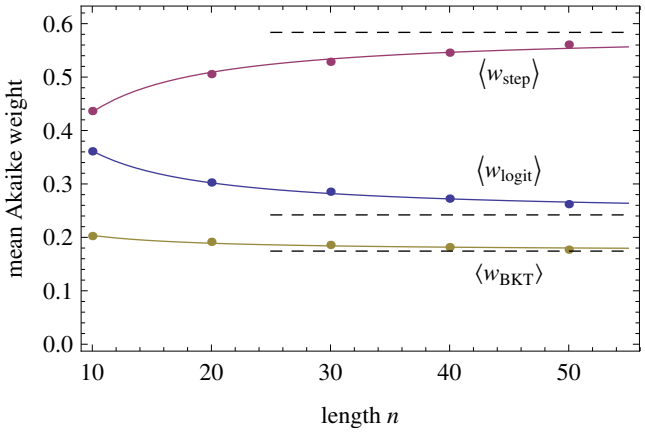

Figure 5: Mean Akaike weights for the three models, P_{step}, P_{logistic}, and P_{BKT}, when fit to randomly generated data of length n. (Each mean is calculated by averaging over 10,000 random bit sequences.) Also shown is a fit to a function of the form $a + b/n$ and a dashed line marking the asymptotic value a. Note that the large differences between the weights persist in the $n \to \infty$ limit.

Our results suggest that the step model may be useful for modeling the learning of an individual student. However, the step model assumes that learning a skill occurs in a single step. Is this how people actually learn? Certainly, everyone has experienced "eureka learning" at some point in their lives. However, it is unclear how well this describes the acquisition of other skills, especially since many KCs are implicit and people are not consciously aware that they even know them [9]. Certainly, if the student performance bit sequence is of the form $00\ldots011\ldots1$, it seems safe to assume that learning occurred all in one step, corresponding to the first 1 in the sequence. However, it is possible that the transition from unmastered to mastery occurs over some number of opportunities and the bit sequence of steps takes on a more complicated form. In a companion paper [14], we introduce a method (based on AIC) that can describe gradual mastery, even though the step model itself assumes all-at-once learning. In that approach, for a given bit sequence, one speaks about the *probability* that learning occurred at a particular step.

Finally, we see that the scatter plot of Akaike weights for student data is remarkably similar to the scatter plots for the random model. This suggests that the student data has a high degree of randomness, and, in general, that study of the random model may be quite useful for better understanding the student data.

4. ACKNOWLEDGMENTS

Funding for this research was provided by the Pittsburgh Science of Learning Center which is funded by the National Science Foundation award No. SBE-0836012.

5. REFERENCES

[1] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Autom. Control*, 19(6):716–723, Dec. 1974.

[2] J. R. Anderson and R. Pelletier. A development system for model-tracing tutors. In L. Birnbaum, editor, *Proceedings of the International Conference of the Learning Sciences*, pages 1–8, Evanston, IL, 1991.

[3] R. S. J. D. Baker, A. B. Goldstein, and N. T. Heffernan. Detecting learning moment-by-moment. *Int. J. Artif. Intell. Ed.*, 21(1-2):5–25, Jan. 2011.

[4] J. Beck and K.-m. Chang. Identifiability: A fundamental problem of student modeling. In C. Conati, K. McCoy, and G. Paliouras, editors, *User Modeling 2007*, volume 4511 of *Lecture Notes in Computer Science*, pages 137–146. Springer Berlin / Heidelberg, 2007.

[5] K. P. Burnham and D. R. Anderson. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. Springer, July 2002.

[6] H. Cen, Kenneth Koedinger, and B. Junker. Learning factors analysis - a general method for cognitive model evaluation and improvement. In *Proceedings of the 8th international conference on Intelligent Tutoring Systems*, pages 164–175, Jhongli, Taiwan, June 2006. Springer-Verlag Berlin, Heidelberg.

[7] M. Chi, K. Koedinger, G. Gordon, P. Jordan, and K. VanLehn. Instructional factors analysis: A cognitive model for multiple instructional interventions. In *Proceedings of the 4th International Conference on Educational Data Mining*, Eindhoven, the Netherlands, June 2011.

[8] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Model. User-Adapt. Interact.*, 4(4):253–278, 1995.

[9] K. R. Koedinger, A. T. Corbett, and C. Perfetti. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Sci.*, 36(5):757âĂŞ798, 2012.

[10] P. I. Pavlik, H. Cen, and K. R. Koedinger. Performance factors analysis âĂŞA new alternative to knowledge tracing. In *Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling*, page 531âĂŞ538, Amsterdam, The Netherlands, The Netherlands, 2009. IOS Press.

[11] G. Schwarz. Estimating the dimension of a model. *Ann. Stat.*, 6(2):461–464, Mar. 1978.

[12] R. Storn and K. Price. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. *J. Global Optim.*, 11(4):341–359, Dec. 1997.

[13] B. van de Sande. Properties of the bayesian knowledge tracing model. Under review, 2012.

[14] B. van de Sande. Measuring the moment of learning with an information-theoretic approach. Under review, 2013.

[15] K. VanLehn. The behavior of tutoring systems. *Int. J. Artif. Intell. Ed.*, 16(3):227–265, Jan. 2006.

[16] K. Vanlehn, P. Jordan, and D. Litman. Developing pedagogically effective tutorial dialogue tactics: Experiments and a testbed. In *Proceedings of SLaTE Workshop on Speech and Language Technology in Education*, pages 17–20, Farmington, Pennsylvania USA, Oct. 2007.

[17] K. Vanlehn, C. Lynch, K. Schulze, J. A. Shapiro, R. Shelby, L. Taylor, D. Treacy, A. Weinstein, and M. Wintersgill. The andes physics tutoring system: Lessons learned. *Int. J. Artif. Intell. Ed.*, 15(3):147–204, Aug. 2005.