BKT Analysis For Big Data documentation

Release 0.9

G-H Gweon and H-S Lee

Jul 01, 2023

CONTENTS

1 Introduction 1

2 Bayesian knowledge tracing (BKT) 3

3 Parameters 5

4 Test analysis 9

5 Other approaches 17

6 References and links 19

7 Indices and Others 21

Bibliography 23

Index 25

CHAPTER

ONE

INTRODUCTION

This document summarizes the idea of using Bayesian analysis for keeping track of student learning.

The model discussed here is the Bayesian knowledge tracking (BKT) model originated by Corbett and Anderson [Corbett1995].

Here, we have our eyes on *big data*. Big data mean *real time* data as much as a large amount of data. It is conceivable that this type of analysis is run in real time and be used to constructively intervene in a student learning process. The Bayesian method looks promising, while it probably is not the only method.

CHAPTER

TWO

BAYESIAN KNOWLEDGE TRACING (BKT)

2.1 Ingredients

Here are some key definitions, basically following [Corbett1995].

knowledge component (KC)
: A skill, a rule, or a piece of principle that a student is supposed to learn through the prepared activity.

lesson
: A sequence of activities, indexed by $n = 1, 2, ...$, whose goal is to teach a KC.

$p(L_n)$
: The probability that the student knows the KC after step $n-1$ and before step n. Defined as such, this is the *prior* probability, as opposed to the *posterior* probability defined below (Eq. (2.2)). The ideal outcome of the lesson is that the series $p(L_n)$ converges to 1. In such a case, the *prior* probability and the *posterior* probability become indistinguishable and define the knowledge state.

Clearly, one must make some estimate of the initial knowledge, $P(L_1)$. Here are the four *parameters* of a BKT model, including $P(L_1)$.

Symbol	Meaning	Definition
$p(L_1)$	Initial knowing	The probability that the student already knows prior to lesson.
$p(T)$	Transition	The probability of becoming knowledgeable at a step.
$p(G)$	Guess	The probability of guessing correctly without knowledge.
$p(S)$	Slip	The probability of make a mistaken choice with knowledge.

Here, all parameters could be assumed to be independent of student, or some or all parameters can be assumed to be dependent on student. For instance, it is reasonable that $p(L_n)$ be taken to be dependent on individual student, while $p(T)$ could be taken as independent of student, if it is largely dependent on the quality of the task.

2.2 Inference chain

The following inference chain is what makes it possible to trace the student knowledge:

$$P(L_1) \to P\left(L_1 | \text{evidence}\right) \to P(L_2) \to P\left(L_2 | \text{evidence}\right) \to \ ... \tag{2.1}$$

The core mechanism of this inference chain is the posterior probability that follows from Bayes' theorem:

$$p\left(L_n | \text{evidence}\right) = \frac{p\left(\text{evidence} | L_n\right) \cdot p(L_n)}{p(\text{evidence})}. \tag{2.2}$$

Here, **evidence** refers to either "getting the correct answer at step n" or "getting an incorrect answer at step n." These two events can be referred to with symbols C_n and I_n, respectively. Their probabilities satisfy the sum rule:

$$P(C_n) + P(I_n) = 1. \tag{2.3}$$

Then, the following two equations follow directly from (2.2) and the definition of $p(S)$:

$$p\left(L_n|\, C_n\right) = \frac{(1-p(S))\cdot p(L_n)}{p(C_n)},$$
$$p\left(L_n|\, I_n\right) = \frac{p(S)\cdot p(L_n)}{1-p(C_n)}.$$

The probability to make the correct choice is given by

$$p(C_n) = p(L_n)\cdot(1-p(S)) + (1-p(L_n))\cdot p(G). \tag{2.4}$$

So the *posterior* probability can be calculated from the *prior* probability assuming that we know what the values of $p(S)$ and $p(G)$.

Now, in order to complete the problem, we must specify how to go from the *posterior* probability at step n to the *prior* probability at step $n+1$. This is where parameter $p(T)$ comes in:

$$p(L_{n+1}) = p\left(L_n|\,\text{evidence}\right) + \left(1 - p\left(L_n|\,\text{evidence}\right)\right)\cdot p(T). \tag{2.5}$$

So, now, one can see that Equations (2.2), (2.3), (2.4), and (2.5) completely specify the Bayesian inference chain for the student knowledge.

2.3 Convergence

In the above, it is clear that the ideal outcome $p(L_n) \to 1$ is a possible point to which the inference chain, Eq. (2.1), can converge to. Near convergence, we get

$$p(C_n) \approx 1 - p(S), \quad p(L_n|\text{evidence}) \approx p(L_n), \quad p(L_{n+1}) \approx p(L_n), \tag{2.6}$$

where in the second expression, "evidence" can be either C_n or I_n. So, when the knowledge is near convergence, the only parameter that determines the pattern of evidence is $p(S)$ according to this model.

2.3.1 For the future

- **How about the perturbation theory near convergence?**
- **Is any other convergence possible?** That is, is there a non-trivial fixed point for the mapping $p(L_n) \to p(L_{n+1})$, regardless of the evidence (in some average sense)? In a simple minded mathematical way, the answer is no. But, it seems possible that the evidence can fluctuate up and down and $p(L_n)$ can stay at the same value on average. Such *semi*-convergence might occur if the student is not really trying, but is randomly choosing answers, out of boredom or fatigue.

CHAPTER

THREE

PARAMETERS

The key part of a BKT application is estimating the best parameters.

The four parameters, $p(L_1), p(G), p(S)$ and $p(T)$, must be optimized in order to ensure the quality of the model. By definition, the best parameters are those that make the model perform the best in predicting the outcome of student work in a task related to the acquired knowledge.

3.1 Slip parameter

This parameter, $p(S)$, is critical, since when $p(L_n)$ approaches 1, this is the single parameter that matters the most, as discussed in Section *Convergence*.

Important as it is, it is also up to some interpretation.

Here is some discussion [Baker2010]:

```
• Recently, there has been work towards contextualizing the guess and slip
  parameters (Baker, Corbett, & Aleven, 2008a, 2008b)

• Do we really think the chance that an incorrect response was a slip is
  equal when

- Student has never gotten action right; spends 78 seconds thinking; answers;
  gets it wrong

- Student has gotten action right 3 times in a row; spends 1.2 seconds
  thinking; answers; gets it wrong
```

Also, in [Gobert2013], two interesting points are made about the slip parameter. First, slip seems to occur more easily for students who initially struggled and then attained the mastery. Second, possibly, the slip parameter is, partially, a reflection of the different student perception of knowledge even when the mastery has been declared by the learning software.

3.1.1 For the future

- In an ideal model, one would think that the slip tendency would decrease as more time is spent to do a task. Perhaps an exponential function of time is appropriate.
- However, one cannot rule out a distraction factor. So, after a certain time, the slip parameter may bottom out, or even go up slightly.
- If the lesson is continued after the mastery is acquired, then $p(L)$ might not be changing much, but $p(S)$ may be decreasing—one might call this a **hardening** of knowledge or a **tempering** of knowledge. It is not enough to acquire knowledge. It is necessary to apply the acquired knowledge to different problems, gain experience, and make it a more *rounded* one. To do this is to reduce $p(S)$ mainly, I think.

These thoughts suggest that $p(S)$ must be made a function of time and the number of activities "during the hardening period." We may regard spending time or trying various applications as **anti-slip hardening process**.

3.2 Optimization

Now, it has been stated already that parameters (four, or more, if the slip parameter is modeled further) must be optimized. In the literature, various approaches seem to have been tried, including a *brute force* approach of making four dimensional grids and evaluating all χ^2 values, and finding the set of parameters that minimize χ^2. This is equivalent to minimizing the "residual* as explained in [BakerWWW].

It appears that many efforts are made on this front—this is understandable since the multi-parameter least squares fit is always a rather ill-defined process due to the local minima. It seems reasonable that this type of approach would work fine, as long as the model is reasonable and the parameter ranges are narrow so that the answers are already clear from the beginning.

In any case, the following approach might be tried as an improvement to a brute force approach by [BakerWWW].

1. Define $D(n)$ as the data to be fit. This is the grade to student response and the value of $D(n)$ within the BKT model is either 1 or 0, in the simple BKT model. However, it could have a value ranging from 0 to 1, end points included.
2. Now, the theory function can be calculated as a function of parameters $T(n; a_i, D)$. This function will give a value ranging from 0 to 1, end points included. Here, a_i's are fit parameters with $i = 1, 2, 3, 4$: they are $p(L_1), p(G), p(S)$ and $p(T)$, respectively.

   > Note that T depends not only n and a_i, but also D, the data itself. So, it is *not* a conventional function, as it is a *functional* of D.—Does the Levenberg-Marquardt *theory* continue to work in this case?

3. Let us assume that the standard Levenberg-Marquardt theory works fine; in fact, we may not even worry about the theory aspect, to some extent, since, well, all we want to achieve is the minimization of χ^2. Then, we can call the Levenberg-Marquardt algorithm for T fitting D, since the algorithm is that of finding the minimum by following the steepest descent.
4. Try random initial values for the initial parameters and make a map of converged results.

3.2.1 For the future

The above approach may be modified to include the *anti-slip-hardening*. In this case, D must be regarded as $D(n,t)$ where t is the short hand notation for all the time information during the lesson.

The fitting procedure above will not change; only the computation of the theory function $T(n;a_i,D,t)$ will be now more complicated. Modeling the hardening process, we should add more input parameters so we will have more than 4 a_i parameters. If we assume that

1. a thought-invoked hardening (the more time student spends, the less slip) is parameterized one time scale parameter, τ,
2. the exercise-driven hardening (the more problems student solves, the less slip) is parameterized by one scale parameter, N,
3. and the threshold for $p(L)$ is given by some number close to 1 (hardening kicks in only if the mastery is nearly achieved),

then we will have seven parameters in total, not four. Within the Levenberg-Marquardt algorithm, this is perfectly doable, while the brute force method will suffer greatly, as the number of parameters increase.

CHAPTER

FOUR

TEST ANALYSIS

The test data from two groups of students were available for testing the idea of a Bayesian analysis.

A python program was written to do the BKT analysis in conjunction with the existing in-house scientific analysis software to see what we can learn from this type of analysis.

4.1 Pre-BKT analysis

Here are two graphs showing the level as a function of time. Two groups, FHSRed10 and bosred1, of students participated in the same types of activity (ramp game). The first group proceeded all the way to the highest level of activities, while the second group did not proceed beyond level 2.

4.2 BKT analysis

4.2.1 Data

For the above data, different levels were defined as different `sets` of data. So, the `fhsred10` group produced four sets of data available for analysis, while the `bosred1` group produced only two.

4.2.2 Symbols

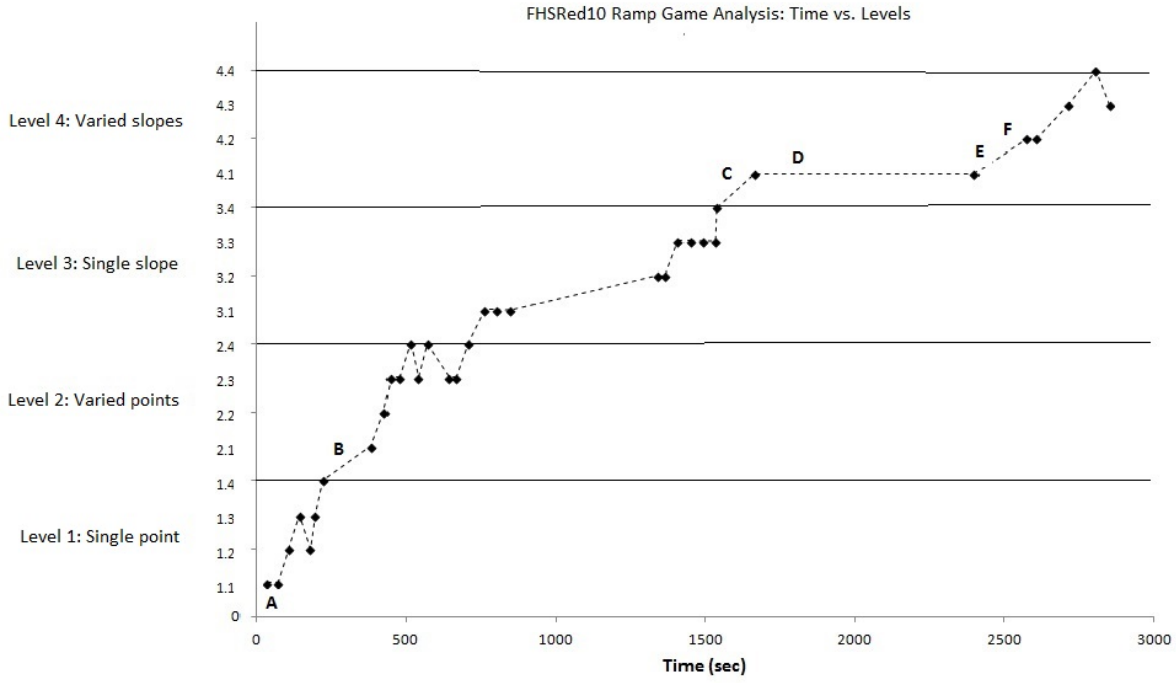
In the program `pg`, `ps`, `pli`, `pt` were used as symbols for $p(G), p(S), p(L_1), p(T)$, respectively. These symbols are explained in Section *Ingredients*.

For each `set` of data, four different initial parameters for the four fit parameters were selected in a completely random manner, each from 0 and 1 (a more restrictive choice is possible [BakerWWW], which will tend to make the results more robust—to do in the future).

4.2.3 Goals of analysis

By carrying out the analysis, we like to obtain numerical estimates of the above four parameters. Also, we like to estimate the real time student knowledge level $p(L)$ by fitting the student score data with the theoretical estimate $p(C)$, the likelihood that the student will get it right. Both $p(L)$ and $p(C)$ are function of the activity index n (cf. *Inference chain*), which is omitted here for brevity.

The activity index n is equivalent to the time, which is used as the x axis in all plots below. In this simple analysis, the time is not used for any other purpose, although perhaps in the future, it could be (cf. *For the future*).

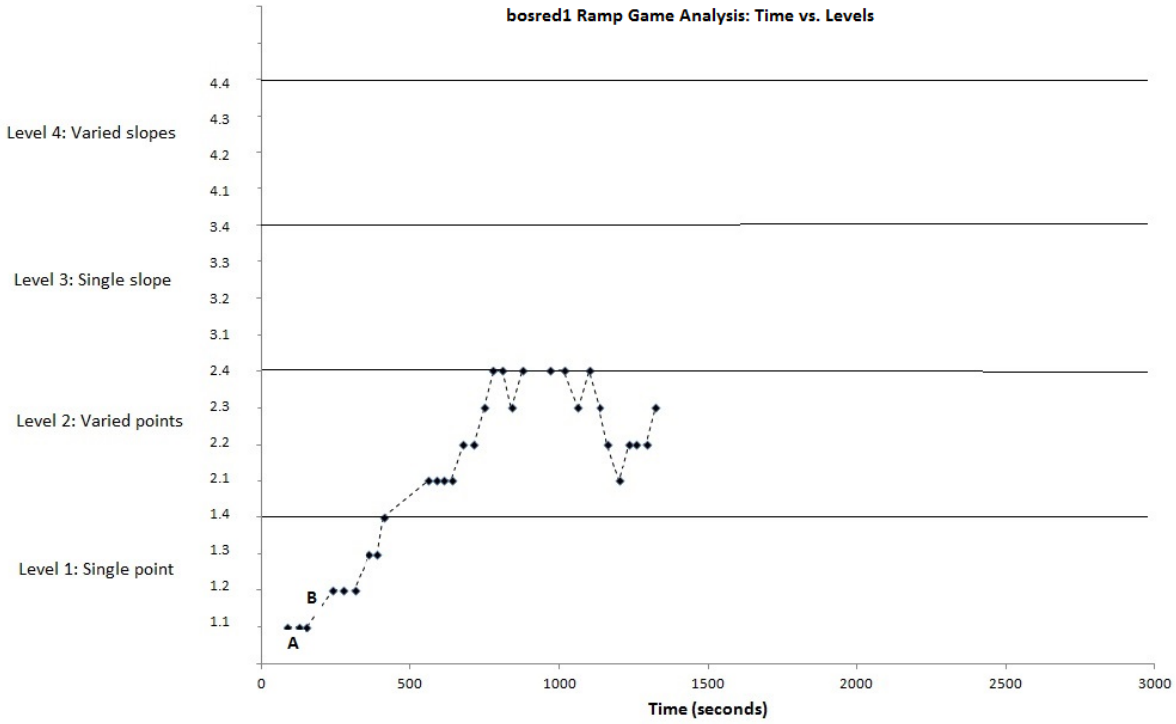


Major Events

A: Table created
B: Graph created: height vs. distance
C: Graph changed: height vs. friction
D: Graph changed: friction vs. distance
E: Graph changed: distance vs. friction
F: Graph changed: friction vs. distance

Time required to complete a level
(measured when students reached 4th sub-level)

Level 1: 188 seconds (7 trials)
Level 2: 322 seconds (10 trials)
Level 3: 774 seconds (10 trials)
Level 4: 1,189+ seconds (6+ trials)



Major Events

A: Table created
B: Graph created: height vs. distance

Time required to complete a level
(measured when students reached 4th sub-level)

Level 1: 324 seconds (9 trials)
Level 2: 764+ seconds (22 and counting trials)

4.2.4 Results for group 1 (fhsred10)

This group completed four different sets of activities. As explained above, four randomly initialized test fits were run on each set. So, four images in a row.

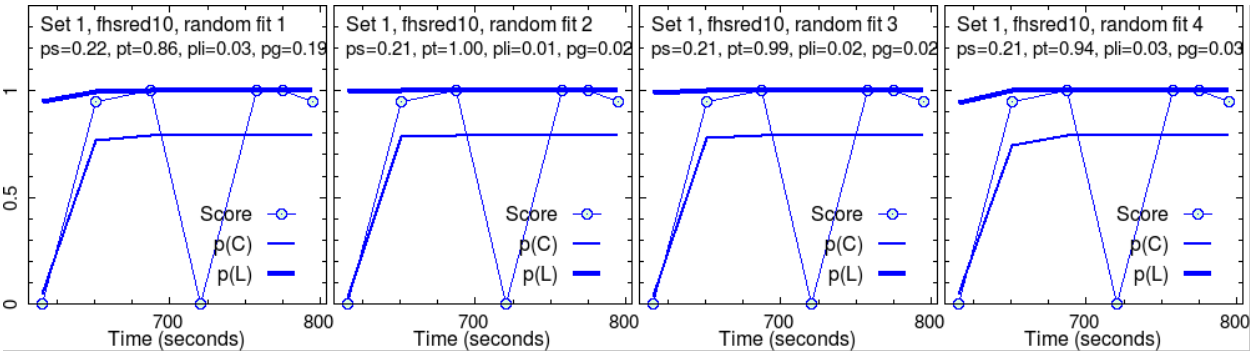
For each randomly initialized fit, the program iteratively searched for the convergence point. Sometimes, this fails. More often, we have success. What we show are successful results only.

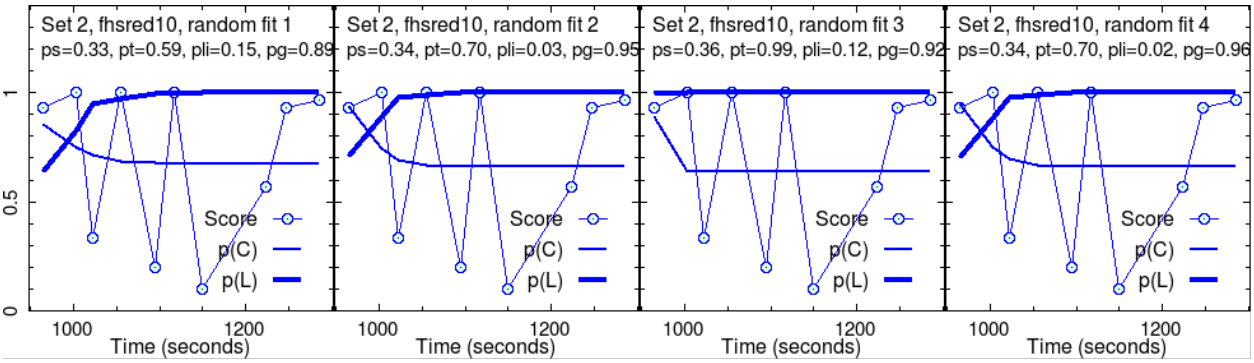
These converged parameters are reported in the second line in each panel. Graphs show the raw score data, along with their fit, $p(C)$, and the estimated knowledge, $p(L)$, that goes with it.

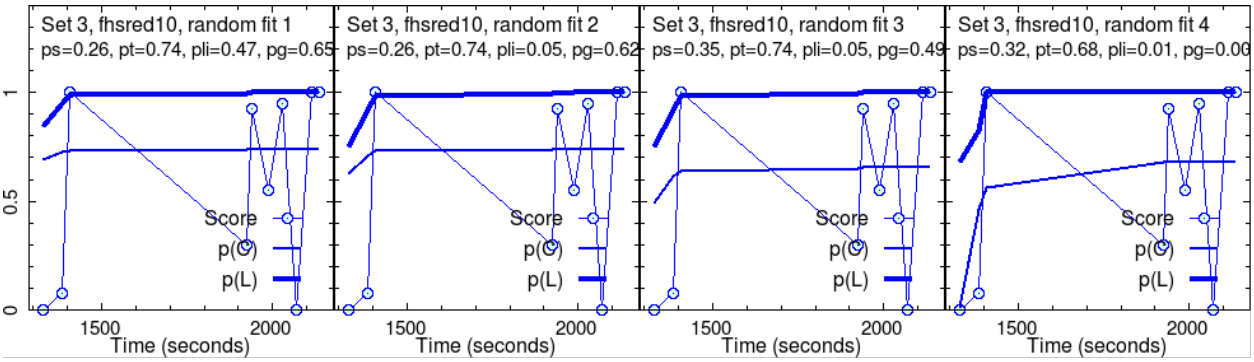
Here, we are concerned with only the robust outcome of the fit: the parameter values corresponding to each fit and the curves of $p(C)$ and $p(L)$.

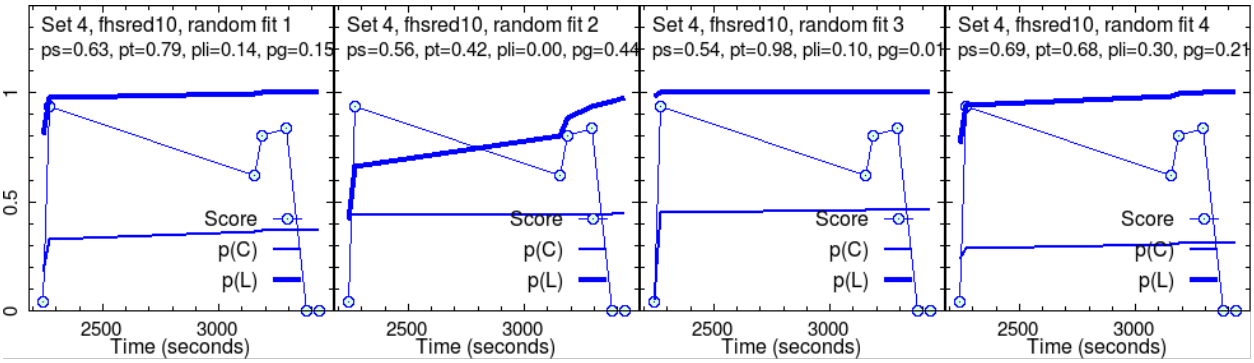
> Side note: The fit program also reports the uncertainties of the converged parameters—however, they are not robust. First, the uncertainty of the data cannot be known for sure, for one thing. Another reason is discussed in Section *Optimization*. Despite this, the *relative* uncertainties between different parameters still make sense. It turns out that by far, $p(S)$ is the most certain parameter that comes out of fits. This is not surprising, given our discussion in Section *Convergence*.

Even if fit results make sense for one fit trial, if the results fluctuate too much between different trials, then it is an indication that those fit results are not reliable. This could be due to an over-modelling (too many fit parameters or too free fit parameters), which may be investigated in the future. Also, more data may help with more statistically meaningful fits.









Some core messages, already

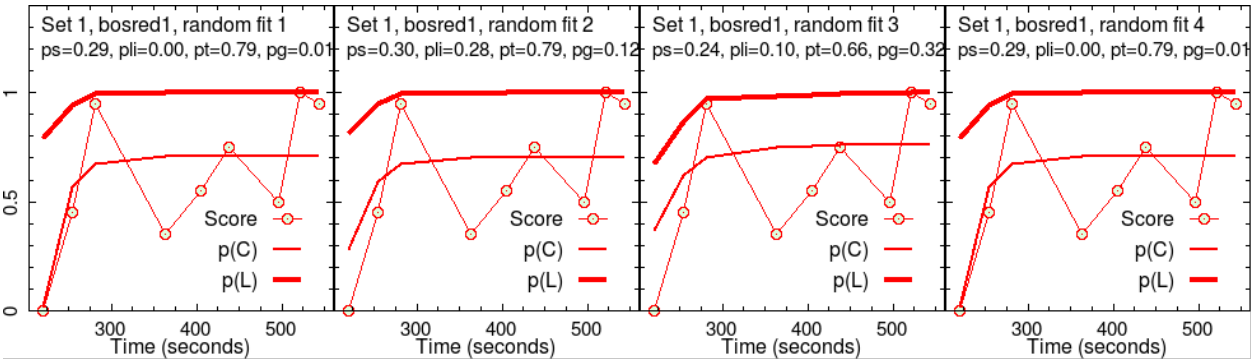
By examining the above results, some core messages appear to form, already.

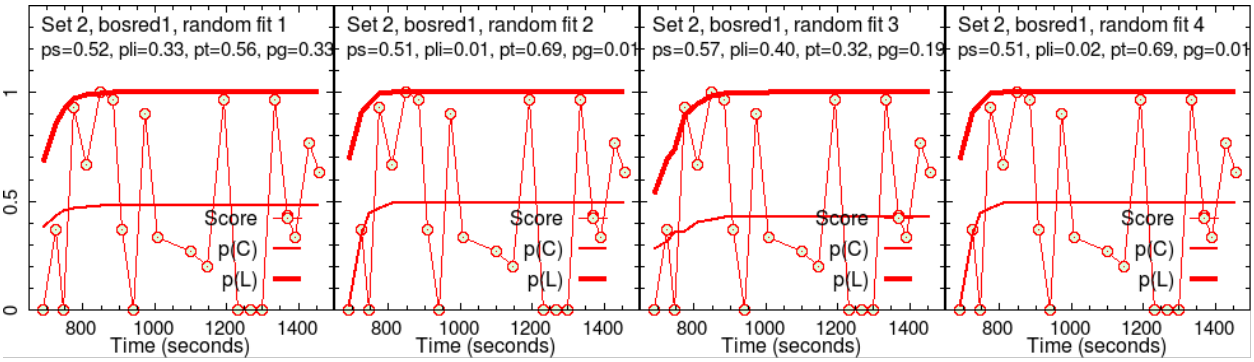
For instance, in set 1, the "effectiveness" of the activities, $p(T)$ is very good, near perfect, $p(T) \approx 1$, for this group. As the level grows, this seems to decrease, in addition with much fluctuation for the most challenging set (set 4).

The slip parameter $p(S)$ is low for set 1, while it becomes quite high for set 4.

4.2.5 Results for group 2 (bosred1)

The second group's results are analyzed the same way.



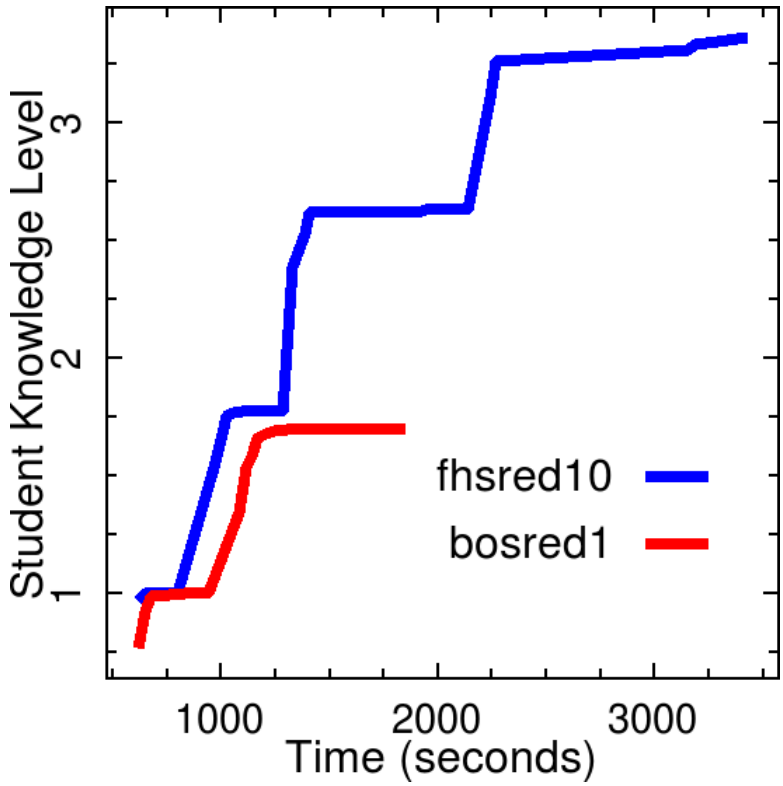


4.2.6 Comparison

It is interesting to see that $p(T)$ for group 2 is much lower than that for group 1, for the same set of activities. Therefore, one would say that the effectiveness of activities was significantly lower fro group 2, which is why they spent more time, doing less.

It is also interesting to see that the slip parameter $p(S)$ is significantly higher for group 2 for set 2 activities. For group 1, a similar level of slipperiness shows up only for set 4.

How about the overall *knowledge level*? Taking the initial value of $p(L)$ for each set as being equal to the final value for the prior set, one can make a *stacked* plot like the following. Here, the first response time for the two group has been taken to be the same.



This graph suggests that group 1 progressed to higher and higher knowledge while group 2 remained at a lower level, showing a slower progression.

As the above discussion makes it clear, $p(L)$ (which is plotted here after stacking and stitching) alone is not reflective

of the learning or the knowledge. It seems that $p(S)$ is also an important parameter that characterizes the student knowledge.

CHAPTER

FIVE

OTHER APPROACHES

I read [Baker2010] that when *multiple skills* are learned in one step, there are other approaches to handle the situation.

- See Conati, Gertner, & VanLehn, 2002; Ayers & Junker, 2006; Pardos, Beck, Ruiz, & Heffernan, 2008.
- However, when there is one primary skill per step, then the BKT is simpler, according to what I read.

CHAPTER

SIX

REFERENCES AND LINKS

http://www.cs.cmu.edu/~listen/BNT-SM/

https://pslcdatashop.web.cmu.edu/

http://www.andestutor.org/

CHAPTER

SEVEN

INDICES AND OTHERS

- genindex
- pdf file
- search

BIBLIOGRAPHY

[Corbett1995] A T Corbett and J R Anderson, *Knowledge Tracing: Modeling the Acquisition of Procedural Knowledge*, User Modeling and User-Adapted Interaction, **4**: 253-278 (1995).

[Baker2010] R S Baker, PSLC Summer School, 2010, Wed-3-BKTPred-v1.pptx (the 4th link of Wedneday).

[Gobert2013] J D Gobert, M S Pedro, J Raziuddin, and R S Baker, *From Log Files to Assessment Metrics: Measuring Students' Science Inquiry Skills Using Educational Data Mining*, J. Learning Sciences, **22**, 521-563 (2013).

[BakerWWW] http://users.wpi.edu/~rsbaker/edmtools.html

INDEX

B

BKT, 1
    Convergence, 4

C

Convergence
    BKT, 4

H

Hardening
    Slip parameter, 5

S

Slip parameter, 5
    Hardening, 5