The Misidentified Identifiability Problem of Bayesian Knowledge Tracing

Shayan Doroudi
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15206
shayand@cs.cmu.edu

Emma Brunskill
Computer Science Department
Stanford University
Stanford, CA 94305
ebrun@cs.stanford.edu

ABSTRACT

In this paper, we investigate two purported problems with Bayesian Knowledge Tracing (BKT), a popular statistical model of student learning: *identifiability* and *semantic model degeneracy*. In 2007, Beck and Chang stated that BKT is susceptible to an *identifiability problem*—various models with different parameters can give rise to the same predictions about student performance. We show that the problem they pointed out was not an identifiability problem, and using an existing result from the identifiability of hidden Markov models, we show that under mild conditions on the parameters, BKT is actually identifiable. In the second part of the paper, we discuss a problem that has been conflated with identifiability, but which actually does arise when fitting BKT models, *semantic model degeneracy*—the model parameters that best fit the data are inconsistent with the conceptual assumptions underlying BKT. We give some intuition for why semantic model degeneracy may arise by showing that BKT models fit to data generated from alternative models of student learning can have semantically degenerate parameters. Finally, we discuss the potential implications of these insights.

Keywords

Bayesian Knowledge Tracing, identifiability, semantic model degeneracy

1. INTRODUCTION

Bayesian Knowledge Tracing (BKT) is a popular model of student learning that tries to predict the probability that a student knows a skill and the probability that a student will answer questions based on the skill correctly. The BKT model is a two state hidden Markov model (HMM) that posits students have either mastered a skill or not, and at every practice opportunity, a student who has not mastered the skill has some chance of attaining mastery. If a student has mastered a skill, they will answer a question correctly unless they "slip" with some (ideally small) probability, and if the student has not mastered the skill, they can only guess correctly with some (ideally small) probability. In 2007, Beck and Chang stated that BKT is not identifiable, meaning that different settings of the four BKT parameters can lead to identical predictions about a student's performance [7]. Whether or not BKT is identifiable is an important issue, because if BKT is not identifiable, it means that we would fundamentally need other criteria (beyond accurately modeling student performance data) to fit BKT models.

However, in this paper, we show that BKT is actually an identifiable model, under mild conditions on the parameters that should always be satisfied in practical settings. This result follows from BKT being a special case of a hidden Markov model and therefore it inherits identifiability results that prior work has proven for HMMs. This implies no additional criteria beyond predictive accuracy are needed to identify a single BKT model that best explains observed student performance, under the assumption that learning can accurately be modeled by a BKT. We then describe three potential issues with BKT models that may have been misconstrued as an identifiability problem in the literature. Note that our goal is by no means to criticize prior researchers, as such researchers helped identify some important limitations of Bayesian Knowledge Tracing, but these limitations do not stem from a lack of identifiablity.

In the second part of this paper, we focus on one of the issues that has been conflated with identifiability, but which actually does arise when fitting BKT models, *semantic model degeneracy*—the model parameters that best fit the data are inconsistent with the conceptual assumptions underlying BKT. We give a critical look at the types of semantic model degeneracy in the literature and then give some intuition for why this problem may arise by showing that BKT models fit to data generated from alternative models of student learning can have degenerate parameters. We further show that fitting models to sequences of different lengths generated from the same underlying model can result in different forms of semantic degeneracy. We show that these insights can have important implications on how these models should be used.

2. BAYESIAN KNOWLEDGE TRACING

The Bayesian Knowledge Tracing model is a two-state hidden Markov model that keeps track of the probability that a student has mastered a particular skill and the probability that the student will be able to answer a question on that skill correctly over time. At each practice opportunity $i \geq 1$ (i.e., when a student has to an answer a question corresponding to the skill), the student has a latent knowledge state $K_i \in \{0, 1\}$. If the knowledge state is 0, the student has not mastered the skill, and if it is 1, then the student has mastered it. The student's answer can either be correct or incorrect: $C_i \in \{0, 1\}$ (where 0 corresponds to incorrect and 1 corresponds to correct). After each practice opportunity, the student is assumed to master the skill with some probability. The BKT model is parametrized by the following four parameters:

- $P(L_0) = P(K_1 = 1)$: the initial probability of knowing the skill (before the student is given any practice opportunities)
- $P(T) = P(K_{i+1} = 1|K_i = 0)$: the probability of mastering a skill at each practice opportunity (if the student has not yet mastered the skill)
- $P(G) = P(C_i = 1|K_i = 0)$: the probability of guessing
- $P(S) = P(C_i = 0|K_i = 1)$: the probability of "slipping" (answering incorrectly despite having mastered the skill)

3. IDENTIFIABILITY

In their 2007 paper, Beck and Chang claimed that BKT is not identifiable, illustrating this with a particular example of three different BKT models [7]. For concreteness we include these models in Table 1. The authors consider the case of predicting the probability of correctness under these three models as the students receive practice opportunities, but in absence of any observation about the student's performance. They use plots as in Figure 1 to claim that the three models make very different predictions about student knowledge (Figure 1 (a)), but make identical predictions about student performance (Figure 1 (b)). They claim,

> All three of the sets of parameters instantiate a knowledge tracing model that fit the observed data equally well; statistically there is no justification for preferring one model over another. This problem of multiple (differing) sets of parameter values that make identical predictions is known as identifiability.

However, this is not correct since no data was used to fit these curves; the curves are predicting the probability that a student will know the skill or will answer the skill correctly at each practice opportunity i, *when we have no prior performance or data on the student.* In order to take past data from a student into account, we actually want to predict $P(K_i = 1|C_1, \ldots C_{i-1})$ and $P(C_i = 1|C_1, \ldots C_{i-1})$ and this is indeed what we do in practice when doing knowledge tracing; we make predictions based on our past observations. Figure 2 shows the curves predicting these conditional probabilities for a particular sequence of correct/incorrect answers for a student (namely we use $(1, 0, 0, 0, 0, 0, 0, 1, 1)$). We find that even when we condition on a single observation (i.e., for $P(C_2 = 1|C_1)$), the three models make vastly different predictions, and as we collect more data, the models continue to make very different predictions. In fact, except for $P(C_1 = 1)$, the models never agree on the probability that a student would answer the step correctly.

Formally, a model is said to be *identifiable* if there are no two distinct sets of model parameters θ and θ' that can give rise to the same joint probability distribution over observations under that model. As far as inference is concerned, identifiability means that the likelihood function of the model has only one global maximum, so inference of the true model parameters is possible. In the case of BKT, the model would be identifiable if for any two distinct sets of BKT parameters, θ and θ',

$$P_\theta(C_1, C_2, \ldots, C_n) \neq P_{\theta'}(C_1, C_2, \ldots, C_n)$$

for some $n \geq 1$. What Beck and Chang show is that there can be infinitely many models that share the same set of marginal distributions $P(C_1), P(C_2), \ldots, P(C_n)$. This does not mean the model is unidentifiable. As we saw from Figure 2, the conditional distribution $P(C_n|C_1, \ldots, C_{n-1})$ is quite different for each model, and so the joint distribution $P(C_1, \ldots, C_n)$ is also very different for the three models.

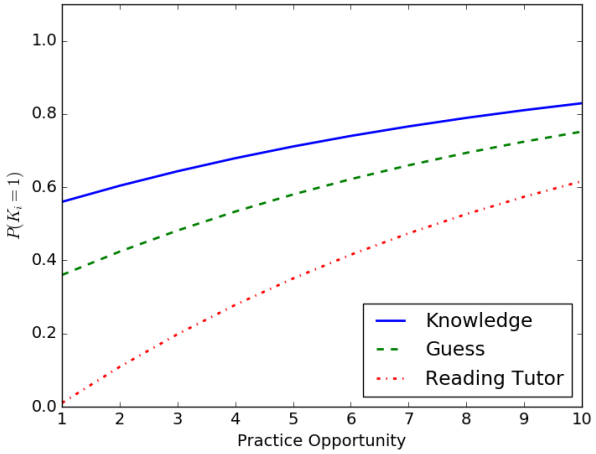
It turns out there has been a substantial amount of work, going back 50 years and continuing to this day, on finding the conditions for which hidden Markov models are identifiable [15, 1, 2, 17, 10]. Although much of the literature focuses on particular types of HMMs (e.g., stationary, irreducible) that do not include the standard BKT model, Anandkumar et al. have recently shown that, subject to some non-degeneracy conditions, a large class of HMMs, which includes BKTs, is identifiable with just the joint probability distributions for up to three sequential observations [4]. That is, knowing $P(C_1), P(C_1, C_2)$, and $P(C_1, C_2, C_3)$ is enough to infer the unique model parameters, subject to non-degeneracy conditions. In our context, the conditions are that $P(L_0) \notin \{0, 1\}$, $P(T) \neq 1$, and $P(G) \neq 1 - P(S)$. This suggests that as long as we have more than two observations per student, BKT models with reasonable parameters are identifiable and there is a single global maximum to the likelihood function. Feng recently independently showed the same result directly for BKT models, except without requiring the condition that $P(L_0) \neq 0$ [9]. One advantage of relying on general identifiability results for HMMs is that we can use the same results to show the conditions under which related student models that can also be modeled as HMMs are identifiable[1].

This misuse of the term "identifiability" has lead to multiple subsequent papers in the educational data mining community throughout the past decade which have similarly given a mistaken description of the underlying phenomena [5, 16, 13, 12]. Two papers, however, have correctly identified that the

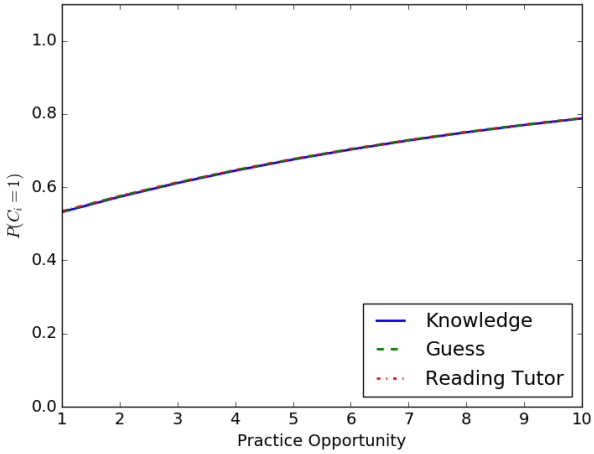
[1]For example, for the BKT model with forgetting, where $P(F) = P(K_{i+1} = 0|K_i = 1) \neq 0$, we can show that the model is identifiable with the same conditions, except that we require $P(T) \neq 1 - P(F)$ instead of $P(T) \neq 1$. We can also easily show the conditions under which multi-state extensions of BKT such as the model introduced in Section 4.2 are identifiable. These conditions can be derived from Condition 3.1 and Proposition 4.2 of [4]. See also the note under Proposition 3.4 of [3].

	Model		
Parameter	Knowledge	Guess	Reading Tutor
$P(L_0)$	0.56	0.36	0.01
$P(T)$	0.1	0.1	0.1
$P(G)$	0	0.3	0.53
$P(S)$	0.05	0.05	0.05

Table 1: The three BKT models used by Beck and Chang [7] to claim BKT is unidentifiable. The models are chosen to have very different semantic interpretations. The Knowledge model requires the student to master the skill to get it correct, the guess model relies on the student guessing, and the Reading Tutor model has an even higher probability of guessing, but it was based on models actually used by the Reading Tutor [14].

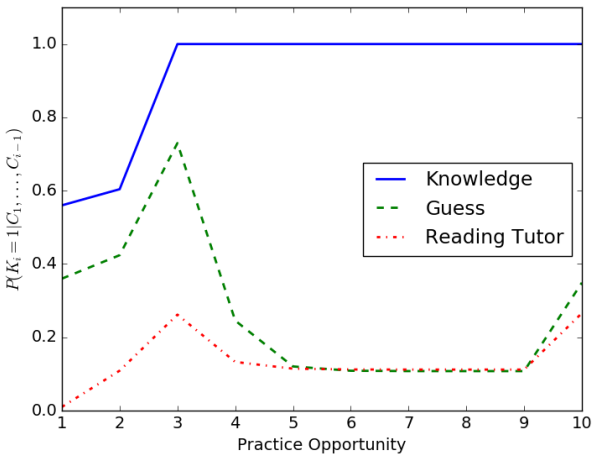


(a) Learning Curve

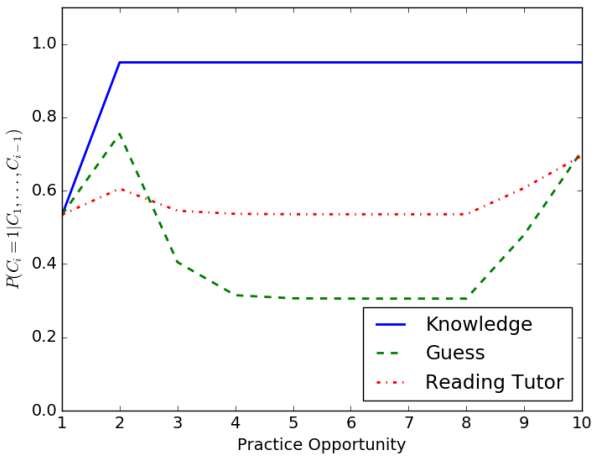


(b) Performance Curve

Figure 1: Hypothetical learning and performance curves for three models from [7], in absence of any data.



(a) Learning Curve



(b) Performance Curve

Figure 2: Learning and performance curves for three models from [7] conditioned on all past observations for a student whose observed trajectory is as follows: $(C_1, C_2, C_3, C_4, C_5, C_6, C_7, C_8, C_9) = (1, 0, 0, 0, 0, 0, 0, 1, 1)$

"identifiability problem" is limited to the case where there is no data [18, 11]. Even though this is not a statistically precise claim, it does show that some researchers have the correct understanding behind the phenomenon. Van de Sande distinguishes between the two cases where predictions are made in the absence of data and where they are made in the presence of data, and claims that the source of the identifiability problem in the former case is that the predictions can be completely determined by three parameters, so there is a degree of freedom [18]. When we are making predictions, however he claims there is no identifiability problem, because $P(K_i|C_i)$ depends on four parameters [18]. While he has correctly identified the absence of an identifiability problem in the presence of data, we believe that there is still confusion about the identifiability problem in the community (e.g., some of the papers that show a misunderstanding of the issue are more recent than [18]). We hope to make the absence of an identifiability problem more clear and elucidate the phenomena and misconceptions surrounding it. Gweon et al. also distinguish between two cases which they refer to as the BKT model without measurement and the BKT model with measurement, and show, as van de Sande did, that the former depends on three parameters (hence the "identifiability problem") whereas the latter depends on all four [11]. However, they claim this does not necessarily mean that the BKT model with measurement does not suffer from an identifiability problem, and actually claim that it still does suffer from an identifiability problem, because empirically, they found that for some data, fitting BKT models many times resulted in a wide spread of possible parameters [11]. However, this cannot be due to the presence of an multiple global maxima, which we have shown cannot exist, and hence must be due to multiple local optima.

The work closest to ours is Feng's recently published dissertation [9]. The author gives a similar explanation to ours for why Beck and Chang's claim was incorrect and also proves that the BKT model is identifiable directly [9]. However, we believe the exposition there is perhaps less accessible to the educational data mining community and will likely not obtain the visibility needed to clear the misunderstandings surrounding the identifiability of BKT. In this paper, we not only focus on identifying the misidentified identifiability problem, but also understanding the confusion surrounding it as well as pointing out actual issues with fitting BKT models that have been conflated with identifiability. This is the focus of the rest of the paper.

There are three potential sources of confusion that we believe could be and have been misconstrued as an identifiability problem:

1. *A priori predictions.* That multiple models, which make very different claims about student's knowledge state over time, could predict the same probability that students answer questions correctly over time *in the absence of data.* This is the problem that Beck and Chang conflated with identifiability, and many researchers thereafter also treated as identifiability. As we showed above, van de Sande, Gweon et al. and Feng correctly identified what is happening here [18, 11, 9].

2. *Multiple local optima.* It is well known that the expectation-maximization algorithm that is commonly used to fit BKT models is suceptible to converging to local optima of the likelihood function rather than converging to the global optimum. While Beck and Chang clearly did not conflate this with the identifiability issue, we saw that other researchers such as Gweon et al. have possibly conflated the two. In order to avoid local optima, one can use a grid search over the entire parameter space or run multiple iterations of the expectation-maximization algorithm with different initializations of the parameters.

3. *Semantic model degeneracy.* Baker et al. identified another problem with BKT models, which they termed model degeneracy [5]. A model is said to be semantically degenerate[2] when it is inconsistent with the conceptual assumptions underlying the BKT model. The problem is when the model that best fits our data is semantically degenerate. Even though Baker et al. clearly contrasted this to the (supposed) identifiability problem, we claim that this is the problem that Beck and Chang attempted to fix in their paper. We will now focus on better understanding this problem.

4. SEMANTIC MODEL DEGENERACY

In their paper, Beck and Chang propose a way to get around the identifiability problem. They propose using Dirichlet priors to encode prior beliefs about the BKT parameters, which will in turn bias the model search towards more reasonable parameters [7]. They motivate their method as follows:

> We have more knowledge about student learning than the data we use to train our models. As cognitive scientists, we have some notion of what learning "looks like." For example, if a model suggest that a skill gets worse with practice, it is likely the problem is with the modeling approach, not that the students are actually getting less knowledgeable. The question is how can we encode these prior beliefs about learning?

The problem they appear to be describing is that some models have parameters that do not match our intuitions of student learning, i.e., they are exactly describing the issue of semantic model degeneracy (and not that of unidentifiability). Baker et al. later provide another solution to tackling semantic model degeneracy by using contextual features to estimate the guess and slip parameters [5]; however, interestingly they did not view Beck and Chang's original solution as a way of tackling semantic model degeneracy, treating it as a way to tackle identifiability as the authors originally claimed.

Having shown that identifiability is not an issue with BKT, and given that there are easy ways to tackle the existence of local optima, we believe semantic model degeneracy is perhaps the most important problem with respect to fitting BKT models that needs to be better understood and tackled. Essentially, the problem arises because the BKT is simply a

[2]We refer to this property as semantic model degeneracy to distinguish it from mathematically degenerate parameters that would result in BKT models being unidentifiable, as described above.

particular form of a two-state hidden Markov model and it will try to fit the best two state hidden Markov model it can to the data; our model fitting procedures do not understand that the $K_i = 1$ state is supposed to correspond to mastering a skill, and so it might fit a model that does not match our intuitions of mastery. We will try to understand this in more detail below, but first we aim to characterize the types of semantic model degeneracy that have been pointed out in the literature.

4.1 Types of Semantic Model Degeneracy

Baker et al. distinguish between two forms of semantic model degeneracy: *theoretical degeneracy* and *empirical degeneracy* [5]. They define a model to be theoretically degenerate when either the guess or the slip parameter is greater than 0.5. They define a model to be empirically degenerate if one of two things occur: (1) for some large enough n the model's estimate of the student having mastered the skill decreases after the student gets the first n skills correct or (2) for some large enough m, the student does not achieve mastery (our estimate of the student having mastered the skill does not go beyond 0.95) even after m consecutive correct responses [5]. The authors arbitrarily chose the values $n = 3$ and $m = 10$. Note that the first form of empirical degeneracy is only possible if $1 - P(S) < P(G)$ (i.e., the student is more likely to answer a question correctly if they have not mastered a skill than if they have mastered a skill), as was shown by van de Sande [18]. This is true, even for $n = 1$. Thus, this first notion of empirical degeneracy is equivalent to $P(G) + P(S) > 1$, which implies either $P(S) > 0.5$ or $P(G) > 0.5$, meaning that it always implies theoretical degeneracy! Huang et al. have noted that while $P(G) + P(S) > 1$ definitely implies semantically degenerate parameters as it contradicts mastery, the condition that $P(G) < 0.5$ and $P(S) < 0.5$ may not always be necessary for the parameters to be semantically meaningful, since, for example, there may be some domains where the student can guess the correct answer easily [12]. We agree that suggesting $P(G) < 0.5$ is degenerate does seem somewhat arbitrary depending on the domain; however, we do think $P(S) > 0.5$ should be characterized as a form semantic degeneracy, because, as Baker et al. claimed, it does not make sense for a student who has mastered a skill to answer questions of that skill incorrectly most of the time—that goes against our intuitions of what mastery means. In any case, it does not seem like the distinction between theoretical and empirical degeneracy is a clear one, so we suggest categorizing the forms of semantic model degeneracy by what they suggest about student learning:

- *Forgetting*: This is a result of $P(G) + P(S) > 1$, which suggests that not only are students not learning, but that students have some probability of losing their knowledge over time. Another way to view this degeneracy is that the state we would conceptually call the mastery state is now the state where performance is worse.
- *Low Performance Mastery*: This is a result of $P(S) > 0.5$. Alternatively, we can set our threshold for low performance mastery to be lower (e.g., P(S) > 0.4).
- *High Performance Guessing*: This is a result of $P(G) > t$, where t is some threshold. As mentioned earlier, this seems like a weak form of degeneracy, as students can often guess an answer easily even if they have not mastered a skill, but we can set t to a large enough value, to make this a form of model degeneracy.
- *High Performance $\not\Rightarrow$ Learning*: This is the second form of empirical degeneracy given by Baker et al. [5]: for some choice of m, the probability that the student has achieved mastery is less than some threshold p (typically taken to be 0.95) after m consecutive correct responses

4.2 Sources of Semantic Model Degeneracy

We will now consider a possible explanation for why BKT models are so prone to semantic model degeneracy (which we believe to be part of the reason that researchers look towards identifiability and local optima to explain the strange parameters that result from fitting BKT models). First of all, note that forgetting degeneracy will occur whenever students actually do forget or when they learn misconceptions; it is not unreasonable to believe that students will sometimes learn and reinforce a misconception, causing their knowledge of some skill to decrease over time. Thus, while this form of degeneracy technically violates our notion of mastery, it is to be expected if we switch the semantic interpretation of the two states and suppose that students forget instead of learn. We now consider sources of the other forms of semantic model degeneracy. We claim that such forms of semantic model degeneracy can result from not accurately being able to capture the complexity of student learning with a two state HMM. When this is the case, fitting the data with a two state HMM will result in trying to find the best fit of the data for a two state HMM, and not to come up with a model that tries to accurately model the data while also matching our intuitions about what it means for a student to have mastered a skill.

To support our claim, suppose student learning is actually governed by a 10-state HMM with ten consecutive states representing different *levels* of mastery. From each state, the student has some probability of transitioning to the next state (slightly increasing in mastery), and from each state, the student has a probability of answering questions correctly, and this probability strictly increases as the student's level of mastery increases. Specifically consider the model presented in Table 2. Now suppose we try to use a standard BKT model to fit data generated from this alternative model of student learning. The first two columns of Table 3 show the parameters of BKT models fit to 500 sequences of 20 practice opportunities or 100 sequences of 200 practice opportunities, both generated from the the model in Table 2. Notice that the model fits (nearly) degenerate parameters in both cases. When we only have 20 observations per student, the model estimates a very high slip parameter; this is because it has to somehow aggregate the different latent states which correspond to different levels of mastery, and since not many students would have reached the highest levels of mastery in 20 steps, it is going to predict that students who have "mastered" the skill are often getting it wrong. However, what's more interesting is that for the same model, if we simply increase the number of observations per student from 20 to 200, we find that the slip parameter is reasonably small, but now the guess probability is 0.49! This is because, by

	State i									
Parameter	0	1	2	3	4	5	6	7	8	9
$P(K_0 = k)$	0.1	0.1	0.1	0.2	0.2	0.3	0	0	0	0
$P(C_i = 1 \mid K_i = k)$	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$P(K_i = k+1 \mid K_i = k)$	0.4	0.3	0.2	0.1	0.05	0.05	0.05	0.05	0.05	-

Table 2: Alternative model of student learning where there are ten levels of mastery.

	10-State HMM		AFM	
Parameter	20	200	20	200
$P(L_0)$	0.30	0.001	0.09	0.001
$P(T)$	0.05	0.02	0.05	0.05
$P(G)$	0.27	0.49	0.14	0.28
$P(S)$	0.44	0.13	0.46	0.03

Table 3: BKT models fit to data generated from the model described in Figure 2 and an additive factors model described in the text. The first column for each model is fit to 500 sequences of 20 practice opportunities, while the second column is fit to 100 sequences of 200 practice opportunities. The models were fit using brute-force grid search over the entire parameter space in 0.01 increments for the parameters using the BKT Brute Force model fitting code [6].

this point most students have actually reached the highest level of mastery, so to compensate for the varying levels of mastery that occurred earlier in student trajectories, the model will have to estimate a high guess parameter. So we find that not only can alternative models of student learning lead to fitting (near) degenerate parameters, but varying the number of observations can lead to different forms of degeneracy! This is a counterintuitive phenomenon that we believe is not the result of not having enough data (students) to fit the models well, but rather the result of the mismatch between the true form of student learning and the model we are using the fit student learning.

We find similar results if we fit a BKT model to data generated from another alternative model of student learning that is commonly used in the educational data mining community, the additive factors model (AFM) [8]. In particular, we used the model

$$P(C_i = 1) = \frac{1}{1 + exp(-\theta + 2 - 0.1i)}$$

where $\theta \sim \mathcal{N}(0, 1)$ is the student's ability[3]. The second two columns of Table 3 show the parameters of BKT models fit to data generated from this model. We again find that when using only data with 20 practice opportunities, we fit a high slip parameter, but when we using data with 200 practice opportunities, we fit a higher guess parameter and a very small slip parameter.

Additionally, notice that for the parameters fit to the 10-state HMM, the probability of transitioning to mastery is very small when we fit to sequences with 200 practice opportunities. Since the transition probability is small and the guess probability is large, we also have high performance $\not\Rightarrow$ learning degeneracy for this model for $m = 10$. That is,

$$P(K_{11} = 1 \mid C_1 = 1, C_2 = 1, \ldots, C_{10} = 1) \approx .89 < 0.95$$

This is yet another form of degeneracy that does not exist in the model fit to sequences of 20 practice opportunities. Furthermore, notice that when we have 200 observations, the probability of transitioning to mastery is smaller than $P(K_i = k + 1 \mid K_i = k)$ for all states i in the model that generated the data (Table 2). Again, this is because the best fitting BKT model will aggregate low performing states and high performing states, so a single transition in the BKT model between these two aggregate states will have to loosely correspond to the student transitioning several times in the actual 10-state HMM. Thus, while the learned BKT model makes it appear as though learning happens very slowly, according to the true student model, learning actually occurs much more often but in more progressive increments. This suggests that if we use some automated technique to detect if a skill is useful for student learning, we may conclude it is not, if we do not allow for the possibility that students are learning progressively.

These observations have important implications for how learned models can be used in practice. Using such a BKT model to predict student mastery can lead to problematic inferences. For example, for the first model in Table 3, the BKT model assumes that when a student has reached mastery, they have a 56% chance of answering a question correctly, whereas a student who has actually mastered the skill will have a 90% chance of answering correctly (see Table 2). Thus, an intelligent tutoring system that uses such a BKT model to determine when a student has had sufficient practice on a problem, will likely give far fewer problems to the student than they actually need in order to reach mastery!

There are several potential ways that future work can proceed in light of these findings. One is that we should be giving our model fitting procedures more domain knowledge about the kind of model we want it to fit. This is essentially what Beck and Chang did by using Dirichlet priors [7] and what Baker et al. did by estimating the guess and slip parameters using context [5]. But perhaps there are other ways of doing this where we do not need to give context-dependent domain knowledge to the model per se, but rather come up with a model that realizes the difference between a student having mastered a skill or not (which the BKT model cannot do). However, this may not be ideal in some cases where student learning cannot accurately be modeled by BKT with semantically plausible parameters. For example when we have forgetting degeneracy, we should probably not force

[3]This model suggests that students who are two standard deviations above the mean initially will answer correctly half the time, and after 20 practice opportunities the average student will answer correctly half the time.

the parameters to suggest learning is occurring when it may not be. Another way to proceed is to consider alternative student models, which is an active area of educational data mining research. Perhaps, obtaining semantically degenerate parameters from a fit should signal that our students may be learning in more complicated ways than the simple BKT model can predict, and so we should try to find alternative models that fit our data better without yielding semantically degenerate parameters. Finally, even if our model is semantically degenerate, it does not necessarily make the BKT model useless. The result of fitting a BKT model is that we get the best fit of the data given that we are modeling the data with a two-state HMM (if we disregard local optima). Presumably, such a model can give us some insights about student learning even if it is not modeling student mastery. So perhaps we can use such semantically degenerate models to understand student learning rather than to predict student mastery.

5. CONCLUSION

We have explored the issues of identifiability and semantic model degeneracy in Bayesian Knowledge Tracing. We have shown that what researchers posited was an identifiability problem is actually not an identifiability problem, and by using a result from the literature on learning hidden Markov models, we showed that an identifiability problem does not exist for BKT models (with the exception of some mathematically degenerate cases that should not come up in practice). We then examined the various issues with fitting BKT models that have been conflated with identifiability. We offered what we believe to be new insights on one potential source of semantic model degeneracy. We believe analyzing the sources of semantic model degeneracy in more detail can be a fruitful direction for future research. For example, it could be useful to know what BKT parameters result from fitting various other popular models of student learning. It would also be informative to see if we can find automated ways of detecting which assumptions of BKT are not met in our data (e.g., the number of levels of mastery, the independence of different skills). Such analyses could help in devising better student models, and ultimately may lead to a better understanding of student learning.

6. ACKNOWLEDGEMENTS

The research reported here was supported, in whole or in part, by the Institute of Education Sciences, U.S. Department of Education, through Grants R305A130215 and R305B150008 to Carnegie Mellon University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Dept. of Education.

7. REFERENCES

[1] E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, pages 3099–3132, 2009.

[2] Y. An, Y. Hu, J. Hopkins, and M. Shum. Identifiability and inference of hidden markov models. Technical report, Technical report, 2013.

[3] A. Anandkumar, R. Ge, D. J. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1):2773–2832, 2014.

[4] A. Anandkumar, D. Hsu, and S. M. Kakade. A method of moments for mixture models and hidden markov models. In *Conference on Learning Theory*, pages 33–1, 2012.

[5] R. S. Baker, A. T. Corbett, and V. Aleven. Improving contextual models of guessing and slipping with a truncated training set. *Human-Computer Interaction Institute*, page 17, 2008.

[6] R. S. Baker, A. T. Corbett, S. M. Gowda, A. Z. Wagner, B. A. MacLaren, L. R. Kauffman, A. P. Mitchell, and S. Giguere. Contextual slip and prediction of student performance after use of an intelligent tutor. In *International Conference on User Modeling, Adaptation, and Personalization*, pages 52–63. Springer, 2010.

[7] J. E. Beck and K.-m. Chang. Identifiability: A fundamental problem of student modeling. In *International Conference on User Modeling*, pages 137–146. Springer, 2007.

[8] H. Cen. *Generalized learning factors analysis: improving cognitive models with machine learning.* Carnegie Mellon University, 2009.

[9] J. Feng. *Essays on learning through practice.* PhD thesis, The University of Chicago, 2017.

[10] É. Gassiat, A. Cleynen, and S. Robin. Inference in finite state space non parametric hidden markov models and applications. *Statistics and Computing*, 26(1-2):61–71, 2016.

[11] G.-H. Gweon, H.-S. Lee, C. Dorsey, R. Tinker, W. Finzer, and D. Damelin. Tracking student progress in a game-like learning environment with a monte carlo bayesian knowledge tracing model. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, pages 166–170. ACM, 2015.

[12] Y. Huang, J. Gonzalez-Brenes, R. Kumar, and P. Brusilovsky. A framework for multifaceted evaluation of student models. In *Proceedings of the 8th International Conference on Educational Data Mining.* University of Pittsburgh, 2015.

[13] J. I. Lee and E. Brunskill. The impact on individualizing student models on necessary practice opportunities. *International Educational Data Mining Society*, 2012.

[14] J. Mostow and G. Aist. Smart machines in education. chapter Evaluating Tutors That Listen: An Overview of Project LISTEN, pages 169–234. MIT Press, Cambridge, MA, USA, 2001.

[15] T. Petrie. Probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 40(1):97–115, 1969.

[16] S. Ritter, T. K. Harris, T. Nixon, D. Dickison, R. C. Murray, and B. Towle. Reducing the knowledge tracing space. *International Working Group on Educational Data Mining*, 2009.

[17] P. Tune, H. X. Nguyen, and M. Roughan. Hidden markov model identifiability via tensors. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 2299–2303. IEEE, 2013.

[18] B. van De Sande. Properties of the bayesian knowledge tracing model. *JEDM-Journal of Educational Data Mining*, 5(2):1–10, 2013.