ABSTRACT
Learning curves are an important tool in cognitive diagnostics modeling to help assess how well students acquire new skills, and to refine and improve knowledge component models. Learning curves are typically obtained from a model estimated on real data obtained from a finite, and usually limited, sample of students. As a consequence, there is some uncertainty associated with estimating the model from that sample, and a risk that the inferences made using learning curves derived from the estimated model are over-confident one way or another. Based on previous work modeling the uncertainty on Additive Factors Model parameters, we derive a principled way to quantify the confidence in learning curves associated with each knowledge component. We show that our approach leads to relatively tight bounds on the learning curves, much tighter than a naive approach relying only on parameter uncertainty. This also reveals a disparity across knowledge components regarding how confident one can be in how well these skills are mastered.

Keywords
Learning Curves, Additive Factors Modeling, Knowledge Cognitive Diagnostics Model

1. INTRODUCTION
Learning curves are a crucial tool for cognitive diagnostics modeling. They help build relevant competency frameworks to accurately measure learners skills and to give them meaningful guidance and feedback in intelligent tutoring systems (ITSs). More precisely, learning curves measure the rate at which students, or simulated artefacts [22], acquire competencies. This allows to evaluate the suitability of a competency framework (aka Q-matrix) and a principled comparison of different learning systems. Learning curves are “graphs that plots performance on a task versus the number of opportunities to practice” [17]. In the educational field, learning curves usually take as learning performance metric the error rate (or equivalently success rate) when applying an individual skill or a set of skills. They were empirically found to follow a “power law of practice” [18], which means that the error rate over time decreases roughly linearly with the logarithm of the number of practice trials taken (aka opportunities). Comparing ITSs or sections of ITSs can be done by considering the steepness of the curve: A steeper curve indicates a faster acquisition of the skills practiced [17].

However, tracking the performance of skills learned in a multidimensional learning environment can be difficult, as those environments combine different set of skills evaluated together. In such situations, some cognitive diagnostic models can be useful to compare learning systems but also to understand the learning mechanisms at play [10]. The Additive Factors Model (AFM) [1], a well known cognitive diagnostics model, does this by assuming that each necessary skill in an item comes with a skill-specific additive contribution towards the probability of success on the item. Fitted AFM parameters can also be used to draw learning curves that compensate for the attrition bias [9]: Over time, fewer learners tend to practice some items because many of them have learned the skill, and the curves tend to quickly degenerate, impacting the estimates of the slopes and the diagnostics of how much learning has occurred. In addition, when learning curves are drawn directly from AFM parameters, the validity of the inferences that can be made will depend greatly on the reliability of the parameters values, and ultimately on the quality of the fitted data. More precisely, fitted parameter values tend to compensate for noise, missing values (e.g. due to attrition) or mis-specified competency models. Rupp and Templin [21] showed for instance how the fitted values of model parameters in DINA [11] would inflate when fitted with purposely erroneous Q-matrices. We can expect a similar impact with any model using Q-matrices, including AFM, a situation made worse by the fact that, in reality, perfect Q-matrices are difficult to identify [5], even when they are retro-engineered from performance data [19]. This motivates the necessity to estimate not only parameter values, but also the statistical confidence on those values, and take into account this uncertainty in any model interpretation, whether based on those values or on the associated learning curves.

Previous work investigated the estimation of standard errors on DINA [20] or AFM [7] parameters, and showed how it could impact learning curves shape and ultimately AFM interpretability and usefulness [15]. Assuming independence across parameters, they produced bounds on learning curves.
using standard confidence intervals on parameter values. However, in practice, the AFM skills parameters (Section 2) are clearly not independent. In this contribution, we show how we can take into account the structure of the covariance between the AFM parameters in order to better model and control the uncertainty on those parameters. We describe a technique for generating confidence intervals on the learning curves using a sampling approach. We illustrate how this works on several competency models from a well-known dataset obtained from a geometry tutoring course, and we show how it allows us to compare how different competency models may model the same skills with different confidence.

In the following Section, we quickly describe the AFM model and introduce our method for obtaining more adequate estimates of the confidence intervals on the learning curves. Section 3 quickly describes the well known EDM dataset that we experiment with in Section 4. Section 5 discusses the results and their impact before we conclude.

2. METHOD
The Additive Factors Model (AFM) introduced by Cen et al. [1, 3] is used in the PSLC-Datashop [12] in order to evaluate domain models. It models the probability of success of a student on item using user and skill specific parameters:

\[ P(Y_{ij} = 1 | \alpha_i, \beta_k, \gamma_k) = \sigma \left( \alpha_i + \sum_{k=1}^{K} \beta_k q_{jk} + \sum_{k=1}^{K} \gamma_k q_{jk} t_{jk} \right) \]

with \( \sigma(x) = 1/(1 + e^{-x}) \) the logistic function, and

- \( \alpha_i \) is the proficiency of student \( i \),
- \( \beta_k \) is the easiness of skill \( k = 1 \ldots K \),
- \( \gamma_k \) is the learning rate for skill \( k \),
- \( Q = [q_{jk}] \) is the \( J \times K \) Q-matrix, representing the cognitive model mapping items to skills,
- \( t_{jk} \) is the number of times student \( i \) has practiced skill \( k \) on any item).

Parameters \( \theta = (\alpha, \beta, \gamma) \) are estimated by maximizing the (penalized) likelihood of the model over observed student outcomes (see e.g. [6]). One attractive feature of AFM is that it easily provides performance curves showing how students acquire skills. Among the different types of learning curves that can be derived from AFM [9, 8], we focus on the data- and student-independent idealized learning curve [8], that simply traces the probability of error for an idealized student with \( \alpha = 0 \) proficiency, on an item with a single skill \( k \):

\[ L_{C_k}(t) = 1 - P(Y = 1 | \alpha = 0, \beta, \gamma) = \sigma (\beta_k + \gamma_k t). \]

Learning curves are typically computed with the maximum penalized likelihood parameters \( \hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\gamma}) \). As noted for example by Philipp et al. [20] and derived for AFM by Durand et al. [7], one can also estimate the uncertainty on \( (\hat{\alpha}, \hat{\beta}, \hat{\gamma}) \), in the form of standard errors. This is relatively straightforward as the covariance matrix on parameter estimates is asymptotically equal to the inverse of the information matrix, \( Cov(\hat{\theta}) = I^{-1}. \) The information matrix \( I_\theta \) can be estimated from first or second order derivatives of the cost function [20, eq. 3, 4]. This also provides a key to quantifying the uncertainty on the learning curves. Using the fact that parameters are (asymptotically) normally distributed around \( \hat{\theta} \) with the known covariance matrix \( Cov(\hat{\theta}) \) [7], we can sample sets of parameters from that multivariate Gaussian distribution, compute the learning curve for each set of parameters, then empirically estimate the error bars on the learning curve through the relevant quantile statistics, as outlined in Algorithm 1.

Although Algorithm 1 focuses on producing error bars on the learning curves, we can also use the simulated sample to evaluate the stability of the entire learning curve, using for example the average standard deviation across opportunities:

\[ \sigma_k = \frac{1}{T} \sum_{t=1}^{T} \text{st.dev.} \{ L_{C_k}(t) \} \]

Lower \( \sigma_k \) indicate that the sampled learning curves are closer together, thus the learning curve is more stable.

3. DATA
For our experiments, we used the “Geometry Area (1996-97)”, a public dataset from DataShop [12]. This dataset contains 6778 observations of the performance obtained by 59 students completing 139 unique items from the “area unit” of the Geometry Cognitive Tutor course (school year 1996-1997). This dataset has been extensively used [1, 2, 7, 13, 14]. We selected three knowledge component (KC) models:

- hLFASearchAICWholeModel3arith0 (referred to simply as arith below),
- hLFASearchModell-context (context below),
- Original (orig below).

These KC models were selected for their reasonable numbers of skills and observations but also because they have distinctive goodness of fit metrics, suggesting that they are high-performing KC models. Table 1 shows that the best predictive model would be arith. The number of skills (KCs) seems to have limited impact on the goodness of fit metrics.  

---

Algorithm 1: Error bars on learning curve for skill \( k \).

**Data:** Parameters \( \theta \), covariance \( Cov(\hat{\theta}) \)

**Parameters:** Target skill \( k \), simulation sample size \( N \)

**Result:** Error bars for the learning curve for skill \( k \), at a set of opportunities (\( t = 1 \ldots T \))

**repeat**

- Sample \( \theta^{(i)} \sim N(\hat{\theta}, Cov(\hat{\theta})) \);
- Compute learning curve \( L_{C_k}^{(i)}(t) \) for target skill \( k \)

**until** \( N \) simulations;

For each opportunity \( t \), compute confidence interval \( \left[ L_{C_k}^{(i)}(t), u_k(t) \right] \) using relevant quantiles of \( \{ L_{C_k}^{(i)}(t) \} \).

---

1. aKa Individual Learning Curve in [9].
Another motivation for choosing these KC models is their skills sharing as some skills have an identical mapping to items in another model, allowing to compare the stability of the same skill across KC models.

4. EXPERIMENTS

In this section, we first illustrate how we derive error bars on the learning curve for a specific KC, then show results for an entire KC model, and finally we compare the stability of learning curves for equivalent skills in different KC models.

4.1 Illustration

We focus on KC#11 (equi-tri-height-from-base/side) from KC-model context. This is a relatively hard ($\beta = -2.97$) skill, but with quick learning ($\gamma = 1.23$). Figure 1 (left) shows the values of $\beta_{11}$ and $\gamma_{11}$ that were sampled by Algorithm 1 for this KC. As seen in the plot, the marginal uncertainty on $\beta_{11}$ and $\gamma_{11}$ is quite high (from -4.5 to -1.5 for $\beta_{11}$), but they are also very correlated: samples with higher easiness have lower learning rate.

Each of the points in Fig. 1 (left) is translated into a corresponding learning curve (Eq. 2) in dotted light gray in Fig. 1 (right). Due to the correlation noted before, we can see that the sampled learning curves are actually fairly stable, compared to what extremes of the distributions of $\beta_{11}$ and $\gamma_{11}$ would suggest (see dashed lines with crosses in Fig. 1, which replicates Fig. 4 from [7]). The red curve in Figure 1 (right) is the learning curve computed from the AFM solution, with 95% confidence intervals obtained from the sample at each opportunity indicated as red bars. We see that although there is some uncertainty around the steep part of the curve, the learning curve is well-controlled and easy to diagnose, indicating that the skill is completely acquired after around 5 opportunities.

4.2 Application to KC models

We now show how we can generate learning curves with confidence intervals for a full KC model. The process illustrated above is applied to each KC, producing one learning curve with confidence bounds. For improved readability, we show the results on KC-model context, which has the smallest number of KCs among our three models.

Figure 2 shows the learning curves for the twelve knowledge components. We can see that most learning curves are well-controlled. The average standard deviation $\sigma$, depending on the skill, ranges from 2% to 8%. “Flat” KCs tend to have lower uncertainty, which is understandable: when the error rate for a skill is low and flat, this is easy for the model to pick up with confidence by predicting high success (high $\beta$) for that skill.

4.3 Comparison of KC models

By better estimating and controlling the uncertainty in learning curves, we can more reliably compare how skills are acquired according to different KC models.

In Figure 3 we show the same skill, compose-by-multiplication, as modeled by the 12-skill model context, and by the 15-skill model orig. The shapes of the learning curves are very similar, which is not surprising as both KCs are associated to the same items, and estimated from the same student outcomes. Despite differences due to the influence of other KCs in the models, the resulting values of $\beta$ and $\gamma$ are similar.

Table 1: Characteristics and predictive quality of the KC models, as computed by PSLC-Datashop.

<table>
<thead>
<tr>
<th>Name</th>
<th>KCs Stud.</th>
<th>#Obs.</th>
<th>AIC</th>
<th>BIC</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>arith</td>
<td>18</td>
<td>59</td>
<td>5104</td>
<td>4948</td>
<td>5569</td>
</tr>
<tr>
<td>context</td>
<td>12</td>
<td>59</td>
<td>5104</td>
<td>5030</td>
<td>5573</td>
</tr>
<tr>
<td>orig</td>
<td>15</td>
<td>59</td>
<td>5104</td>
<td>5180</td>
<td>5762</td>
</tr>
</tbody>
</table>

Figure 1: Left: Sampled $\beta$ and $\gamma$ for KC#11 of the context model. Right: Corresponding learning curves (in light gray); the LC given by the AFM model is in red, with 95% confidence intervals at opportunities up to 10 shown as red vertical bars. The 95% CI from [7] is indicated in black crosses for comparison.
The error bars, however, show that the confidence is slightly better in the orig model, showing an average dispersion of around 3.5% error across the learning curve (versus 4.3% in context). This shows that even in a model with more KCs, learning curves can be modelled with higher confidence.

Our second example, in Figure 4, compares similar skills, compose-subtract from arith, and Subtract from orig. Again, the general shape of the learning curves are similar, due to similar values for the estimated $\beta$ and $\gamma$ in each model. The sampled learning curves also seem quite similar, sug-

For arith, $\beta = .588 \pm .524$ and $\gamma = .329 \pm .200$, while for
Figure 4: KC compose-subtract from model arith (left) and KC Subtract from orig (right). $\bar{\sigma}$ is the average uncertainty across opportunities (lower is better).

Figure 5: Structure of the correlation between $\beta$ (y-axis) and $\gamma$ (x-axis) for all KCs in model context.

One straightforward outcome of this work is that the proposed method provides a much better estimate of the confidence in a learning curve than the method proposed in [7], which relied on the marginal distribution of AFM parameters $\beta$ and $\gamma$ and used the boundaries of straight confidence intervals on each parameter independently. We included their 95% confidence interval as black crosses in Figure 1: that suggests that the uncertainty on the learning curve is high up to 8 or more opportunities. By contrast, our approach shows that the actual uncertainty is much better controlled, and that the skill is essentially learned by opportunity 5 or 6.

In this paper, we have worked with the basic learning curve called the individual learning curve in [9] or the idealized learning curve in [8]. We note that this work can be applied to any learning curve that relies on the parameters of the AFM model. This includes in particular the completed learning curve [9], where empirical observations of success/failure are completed by model estimates.

In previous work, Harpstead and Aleven [10] used empirical learning curve analysis to inform educational game design. They derive empirical curves and AFM-fitted curves, with standard errors on the curves, using a completely different approach from ours. Contrary to the approach advocated here, which relies on the core uncertainty on model parameters resulting from a maximum (penalized) likelihood estimation, their learning curves and error bars are obtained using non-parametric smoothing (LOESS [4], presumably from the stat-smooth function of the ggplot2 R package). On the empirical measurements of success, this produces learning curves that are based on observations alone, and therefore may not have the desirable properties enforced by the AFM model, such as monotonicity (decreasing learning curves). On the fitted AFM predictions, those properties are enforced and apparent from the learning curves.\(^\text{4}\) Two key differences with our approach, however, are:

1. The use of fitted AFM values to produce error rate

\(^{4}\text{Blue curves in [10], Figs 3, 4 and 7.}\)
predictions does not take into account the uncertainty in parameter values due to estimation from a finite sample, and

2. The width of the error bars are directly impacted by the number of students at each opportunity, typically resulting in widening error bars as attrition kicks in. By contrast our sampling-based algorithm often yields narrowing error bars as opportunities increase and the error rates near zero (for all sampled parameters).

A more systematic study of differences between our approach and the non-parametric smoothing of model estimates would require further study. The opportunity of combining both approaches in order to take into account the uncertainty due to parameter estimation and sampling uncertainty across the finite set of students seems particularly promising.

6. CONCLUSION
In this contribution, we provided a principled way to estimate and control the confidence in learning curves derived from the Additive Factors Model. Error bars on the learning curves account for the statistical uncertainty associated with estimating the AFM model from a finite set of students. They allow to more accurately and more confidently interpret how skills are acquired by students. We showed how this allows to characterize learning for all skills of a KC model of a geometry tutoring course. We also showed how modeling the confidence of learning curves can help compare how two different KC models represent the same skill. Our approach was illustrated here on one type of learning curve, but it can be applied to any alternative learning curve, as long as it can be computed from the usual AFM parameters. In addition, the same idea can be applied in a straightforward way to any cognitive diagnostic model for which a covariance on parameters can be computed. This includes in particular, models estimated by penalized maximum likelihood. For instance, the Individualized-slope Additive Factors Model (IAFM) [16], that extends AFM with a student learning rate, could be an excellent candidate to our method, especially as authors noticed that IAFM "[student] learning rate is significantly related to estimates of student ability". Finally, our hope is that this work will help spread the use of learning curves with well-controlled confidence among practitioners of AFM.

7. REFERENCES


