Towards reliable and valid measurement of individualized student parameters

Ran Liu
Human-Computer Interaction Institute
Carnegie Mellon University
ranliu@cmu.edu

Kenneth R. Koedinger
Human-Computer Interaction Institute
Carnegie Mellon University
koedinger@cmu.edu

ABSTRACT

Research in Educational Data Mining could benefit from greater efforts to ensure that models yield reliable, valid, and interpretable parameter estimates. These efforts have especially been lacking for individualized student-parameter models. We collected two datasets from a sizable student population with excellent "depth" – that is, many observations for each skill for each student. We fit two models, the Individualized-slope Additive Factors Model (iAFM) and Individualized Bayesian Knowledge Tracing (iBKT), both of which individualize for student ability and student learning rate. Estimates of student ability were reliable and valid: they were consistent across both models and across both datasets, and they significantly predicted out-of-tutor pretest data. In one of the datasets, estimates of student learning *rate* were reliable and valid: consistent across models and significantly predictive of pretest-posttest gains. This is the first demonstration that statistical models of data resulting from students' use of learning technology can produce reliable and valid estimates of individual student learning rates. Further, we sought to interpret and understand what differentiates a student with a high estimated learning rate from a student with a low one. We found that learning rate is significantly related to estimates of student ability (prior knowledge) and self-reported measures of diligence. Finally, we suggest a variety of possible applications of models with reliable estimates of individualized student parameters, including a more novel, straightforward way of identifying wheel spinning.

Keywords

Explanatory models, model interpretability, individualized parameters, 3, Additive Factors Model, individualized Bayesian Knowledge Tracing

1. INTRODUCTION

In Educational Data Mining, statistical models are typically evaluated based on fit to overall data and/or predictive accuracy on test data. While this is an important initial step in evaluating the contributions of advancements in statistical and cognitive modeling, research in the field could benefit from greater efforts to ensure that models are reliable and valid. More reliable and valid models offer more explanatory power, contributing to the advancement of learning science. They also inspire greater confidence that deploying model advancements in future tutoring systems will genuine result in the hypothesized improvements to learning.

Some recent work has been done towards interpreting, validating, and acting upon cognitive/skill modeling improvements [7, 8, 10, 11, 17]. Educational data mining efforts oriented around personalizing student constructs [3, 12, 13, 14, 18], however, have remained focused on improving predictive accuracy and/or demonstrating hypothetical time savings. Little has been done to validate or understand the estimates that models with individualized or clustered student parameters produce. Anecdotally, efforts to do so have shown that these individualized student parameter estimates, or discovered student clusters, are often difficult to interpret.

It is especially critical to examine the reliability and validity of parameter estimates for modeling advancements that dramatically increase the parameter count, as is generally true for individualized student-parameter models. More parameters create greater degrees of freedom and increase the likelihood that the model may be underdetermined by the data.

We focus on the question: To what degree can we trust a model's parameter estimates to correctly represent the constructs they are supposed to?

Key to expecting reliable, valid estimates of student-level constructs is not just big data in the "long" sense, but big data in the "deep" sense. Oftentimes, the datasets used in secondary analyses in EDM are large in terms of total number of students (or total observations) but highly sparse in terms of observations per skill, per student. These features make it difficult to get reliable measurements of constructs at the individual student level, particularly constructs related to learning over time.

Here, we collected two datasets from a sizable student population (196 students) with excellent "depth" – that is, many observations for each skill for each student. We then fit two models that individualize for student ability and student learning rate (the Individualized-slope Additive Factors Model [9] and Individualized Bayesian Knowledge Tracing [18]). We assess the models' fit to data and predictive accuracy. We also move beyond these metrics to examine the reliability of the models' estimates of student ability and student learning rate. Additionally, we externally validate the parameter estimates against out-of-tutor assessment data.

We further interpret and understand the constructs by visualizing representative student learning trajectories, examining the relationship between estimated student ability and student learning rate, and the relationship between those constructs and self-reported data on motivational attributes. Finally, we propose some useful applications of reliable and valid individualized student-parameter models, including a new way to detect wheel spinning.

2. PRIOR WORK

Prior work on individualizing student parameters has focused on variants of Bayesian Knowledge Tracing (BKT) [3]. This work includes modeling the parameters separately for each individual student instead of separately for each skill [3], individualizing the P(Init) ("initial knowledge") parameter for each student [13], and individualizing both P(Init) and P(Learn) ("learning rate") to the

base BKT model [18]. These models have generally focused on assessing predictive accuracy improvements relative to their respective non-individualized baseline models.

There have also been some "time savings" analyses [12, 18] that evaluate the hypothetical real world impact that individualizing statistical model fits could have. These analyses report the effect of fitting individualized BKT models, compared to traditional BKT, on the hypothetical number of under- and over- practice attempts that would be predicted for each student. Results generally have indicated that many more practice opportunities are needed for models to infer the same level of knowledge when using whole-population parameters rather than individual student parameters. These analyses show that individualized models differ in their hypothetical decision points if they were to be applied to drive mastery-based learning, but they do not in and of themselves interpret the individualized parameter estimates, nor do they assess the reliability and validity of such estimates.

In a previous effort to better understand individualized student learning rate parameters [9], we examined predictive accuracy and parameter reliability in an extension of the Additive Factors Model [2] applied to existing educational datasets. We did not find evidence that individualizing student rate parameters consistently improved predictive accuracy improvements, nor could we validate the parameter estimates on out-of-tutor assessment data. However, the datasets we analyzed either contained a small number of students or were largely sparse in observations for student-skill pairs, with the exception of two datasets. These two datasets happened to be the ones on which the Individualized-slope Additive Factors Model *did* achieve higher predictive accuracy. Thus, we wondered if the sparsity of the datasets were the primary limitation, rather than the modeling advancement itself. This idea is corroborated by the fact that pooling students into "groups" rather than generating individualized estimates worked well on those datasets [9].

For the present modeling work, we collected our own data in order to ensure the data features that we believe are necessary for reliable, valid, and potentially meaningful estimates of constructs at the individual student level.

3. METHODS

It is common in EDM to do secondary analyses across multiple datasets. However, it can be difficult to find datasets that (1) contain a sizable number of students, (2) contain many observations for each skill for each student (i.e., are not sparse), (3) contain students spanning a range of abilities in the domain covered by the tutor, and (4) contain data from out-of-tutor assessment data that is well-mapped to the content in the tutor.

For the present work, we wanted to use as close to an "ideal" dataset as possible for estimating student parameters. We collected our own dataset with a sizable number of students (196), many observations (5-50, depending on the skill) for each skill for each student. In addition, we ensured that a wide range of student ability levels was represented in our data to allow for the possibility that models could capture this variability.

3.1 Data Collection

196 students, spanning 10 classes taught by three different teachers, enrolled in high school geometry participated in two studies conducted about a month apart. A range of student abilities were included in the study. Two of the 10 classes were "Honors" and three of the 10 classes were "Inclusion". Honors classrooms are intended for students who have strong theoretical interests and abilities in mathematics. Inclusion classrooms are "general education" classrooms designed to provide the opportunity for individuals with disabilities and special needs to learn alongside their non-disabled peers.

Students spent five consecutive days participating in each study during their regular geometry class periods. On the first and last days, they took a computerized pretest and posttest, respectively. During the middle three days, they worked within an intelligent tutoring system [19] designed to give them practice on their current chapter's content. This procedure applied to both studies, one of which covered the students' Chapter 3 content (Parallel Lines Cut by a Transversal, Angles & Parallel Lines, Finding Slopes of Lines, Slope-Intercept Form, Point-Slope Form) and the other of which covered the students' Chapter 4 content (Classifying Triangles, Finding Measures of Triangle Sides & Angles, Triangle Congruence Properties). Figure 1 shows an example problem interface from the intelligent tutoring system, which was designed using Cognitive Tutor Authoring Tools [1].

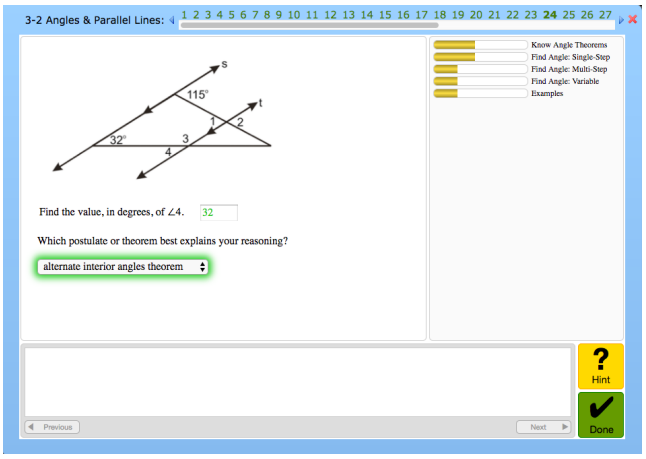

Figure 1. Example problem interface from the intelligent tutoring system used for data collection.

We also collected self-report survey data on motivational factors falling along three dimensions. These were Competitiveness (e.g., "In this unit, I am striving to do well compared to other students" and "In this unit, I am striving to avoid performing worse than others"), Effort (e.g., "I am striving to understand the content of this unit as thoroughly as possible" and "I work hard to do well in this class even if I don't like what we are doing"), and Diligence (e.g., "when class work is difficult, I give up or only study the easy parts" [inverted scale] and "I am diligent"). Self-report measures were indicated on a Likert scale from 1-7.

A key reason we collected two datasets, covering two distinct chapters of the curriculum, is that we were interested in investigating the consistency of student-level parameter estimates across different content, time, and contexts. We discuss this further, along with preliminary results, in Section 4.4.1.

3.2 Statistical Models

3.2.1 The Individualized-slope Additive Factors Model (iAFM)

The Additive Factors Model (AFM) [2] is a logistic regression model that extends item response theory by incorporating a growth or learning term.

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \theta_i + \sum_{k \in KCs} Q_{jk}(\beta_k + \gamma_k T_{ik}) \quad (1)$$

This statistical model (Equation 1) gives the probability p_{ij} that a student i will get a problem step j correct based on the student's baseline ability (θ_i), the baseline easiness (β_k) of the required knowledge components on that problem step (Q_{jk}), and the improvement (γ_k) in each required knowledge component (KC) with each additional practice opportunity. This KC slope, or "learning rate," parameter is multiplied by the number of practice opportunities (T_{ik}) the student already had on it. Knowledge components (KCs) are the underlying facts, skills, and concepts required to solve problems [6].

Individualized-slope AFM (iAFM) builds upon this baseline model by adding a per-student learning rate parameter (δ_i). This parameter represents the improvement (δ_i) by student i with every additional practice opportunity with the KCs required on problem step j.

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \theta_i + \sum_{k \in KCs} Q_{jk}(\beta_k + \gamma_k T_{ik} + \delta_i T_{ik}) \quad (2)$$

The KC and student learning rate parameters are both multiplied by the number of opportunities (T_{ik}) the student already had to practice that KC.

3.2.2 Individualized Bayesian Knowledge Tracing (iBKT)

Bayesian Knowledge Tracing (BKT [3]) is an algorithm that models student knowledge as a latent variable using a Hidden Markov Model. The goal of BKT is to infer, for each skill, whether a student has mastered it or not based on his/her sequence of performance on items requiring that skill. It assumes a two-state learning model whereby each skill is either *known* or *unknown*. There are four parameters that are estimated in a BKT model: the initial probability of knowing a skill a priori – p(Init), the probability of a skill transitioning from not known to known state after an opportunity to practice it – p(Learn), the probability of slipping when applying a known skill – p(Slip), and the probability of correctly guessing without knowing the required skill – p(Guess). Fitting BKT produces estimates for each of these four parameters for every skill in a given dataset. BKT models are usually fit using the expectation maximization method (EM), Conjugate Gradient Search, or discretized brute-force search.

Individualized Bayesian Knowledge Tracing (iBKT [18]) builds upon this baseline BKT model by individualizing the estimate of the probability of initially knowing a skill, p(Init), and the transition probability, p(Learn), for each student. To accomplish the student-level individualization of these parameters, each of them is split into skill- and student-based components that are summed and passed through a logistic transform to yield the final parameter estimate. Details on the decomposition of p(Init) and p(Learn) into skill- and student-based components are described in [18].

4. RESULTS

4.1 Model Fit & Predictive Accuracy

As a first pass evaluation of the two individualized models, we assessed them using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), which are standard metrics for model comparison, and 10 independent runs of split-halves cross validation (CV). Although 10-fold cross validation has been popular in the field, [4] showed that it has a high type-I error due to high overlap among training sets and recommended at least 5 replications of 2-fold CV instead.

Here, the comparison of interest is each individualized model against its non-individualized counterpart. We do not encourage a literal comparison between the predictive accuracies of the two classes of models due to differences in whether they use incoming test data towards their predictions on later test data (BKT/iBKT do, and AFM/iAFM do not).

Both iAFM and iBKT outperform their non-individualized counterparts by all metrics, with the exception of BKT having a better BIC value than iBKT for the Chapter 4 dataset. This is not surprising, as BIC is known to over-penalize for added parameters. We recommend cross validation as a better indicator that iBKT is the true better fitting model in this case.

Counter to the majority of findings reported in [9], iAFM achieved higher predictive accuracy than AFM in both datasets here. This further supports the idea that the "depth" of the dataset is a critical factor in whether an individualized student-parameter model can explain unique variance in the data.

Table 1. Summary of Model Fit and Predictive Accuracy metrics comparing AFM vs. iAFM and BKT vs. iBKT. Cross-validation values are mean RMSE values across 10 runs, with standard deviations included in parentheses.

Data Set	Model	AIC	BIC	CV *Test* RMSE (10-Run Average)
Ch. 3	AFM	57229	57283	0.38440 (0.0039)
	iAFM	**55931**	**56003**	**0.37868 (0.0044)**
	BKT	66714	67473	0.4222 (0.0005)
	iBKT	**56325**	**60479**	**0.3777 (0.0006)**
Ch. 4	AFM	18059	18106	0.41037 (0.0048)
	iAFM	**17863**	**17925**	**0.40789 (0.0050)**
	BKT	19908	**20376**	0.44091 (0.0014)
	iBKT	**18285**	21809	**0.40725 (0.0018)**

4.2 Reliability of Student Parameters

Next, we examined the degree to which we can rely on these parameters to reasonably estimate the constructs that they should be estimating. We believe that a strong relationship between the parameter estimates of two statistical models with entirely different architectures is a high bar for testing reliability. That is, if a student genuinely displayed evidence of high overall ability in a dataset (relative to his/her peers), then both iAFM and iBKT should estimate that to be the case.

Because of known and observed nonlinear relationships between logistic regression and Bayesian Knowledge Tracing parameter estimates, we measured correlation based on Spearman's coefficient (r_s), which is based on rank order.

We observed strong and statistically significant correlations between iAFM Student Intercept and iBKT Student p(Init) parameter estimates (Figure 2, top row). We also observed a strong and statistically significant correlation between iAFM Student Slope and iBKT Student p(Learn) parameter estimates for one of the two datasets (Chapter 4). This correlation was much milder, though still significant, for the other dataset (Chapter 3).

We hypothesize that this difference between datasets may be due to the presence of more difficult KCs in Chapter 4. A dataset with more difficult items should provide more sensitive measures of individual differences in improvement, since it avoids ceiling effects. Indeed, this was the case: the mean KC easiness parameter estimate (β_k) for chapter 4 was 0.799 (which translates to a

probability of 0.69), compared to 1.253 for chapter 3 (which translates to a probability of 0.78). When students are practicing many opportunities at ceiling (which was the case in particular for chapter 3, based on exploratory analyses of the data), the individualized models will often assign them a lower "learning rate" due to an essentially flat learning trajectory.

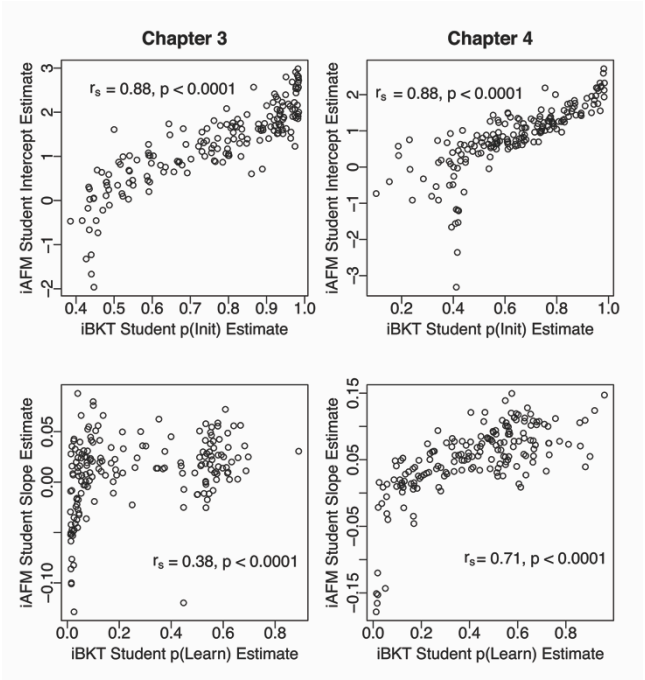

Figure 2. Relationships between iAFM Student Intercept and iBKT Student p(Init) parameter estimates (top row), and between iAFM Student Slope and iBKT Student p(Learn) parameter estimates (bottom row), for the two datasets.

4.3 Validity of Student Parameters

To assess the validity of student parameter estimates, we related them to out-of-tutor assessments of the relevant student constructs. In this case, we validated parameter estimates using pretest and posttest assessment data collected in the study.

4.3.1 Estimates of Student Ability

The Student Intercept (θ_i) parameter of iAFM and the Student p(Init) parameter of BKT are designed to estimate baseline student ability, as least for the knowledge domain represented in the dataset. To validate the models' estimates of this construct, we examined relationships between the model estimates and students' pretest scores, which are an out-of-tutor assessment of student initial ability for the skills covered by the tutor.

We report standard Pearson correlation coefficients here, since the relationships between pretest scores and the parameter estimates did not appear to be particularly nonlinear.

Figure 3 illustrates a summary of these relationships. Both models' estimates of the student ability construct were strongly and significantly correlated with pretest scores.

In addition, adding an individualized student slope *improved* the validity of the model's estimate of student ability (a parameter that's modeled in both AFM and iAFM). We compared the correlations between AFM's intercept estimates to pretest scores (Chapter 3: r = 0.62, p < 0.0001, Chapter 4: r = 0.58, p < 0.0001) to iAFM's intercept estimate / pretest score correlations (Chapter 3: 0.74, p < 0.0001, Chapter 4: r = 0.66, p < 0.0001).

This has several interesting implications for educational applications. First, it suggests that formative assessment via modeling of process data as learning unfolds is a reasonable method of assessment.

It also suggests that detailed assessment data (e.g., from a pretest) could be used to reasonable effect to improve different students' "on-line" estimates of students' knowledge of KCs. For example, combining KC parameter estimates (derived from model-fitting to prior domain-relevant data) with student intercept priors based on pretest assessment data would allow a model like AFM to generate individualized predictions of how much each student needs to practice to reach mastery.

In addition, these results suggest that individualized BKT models could use pretest assessment data to "set" reasonably valid student-specific p(Init) values before collecting any within-tutor data from those students.

In considering the degree to which these results may generalize, it is important to note that the pretests in the present datasets were specifically designed to map closely to the practice problems in the intelligent tutor. Pretests contained 1-2 questions for each KC that was practiced in the tutor, and the items were similar to those encountered within the tutor.

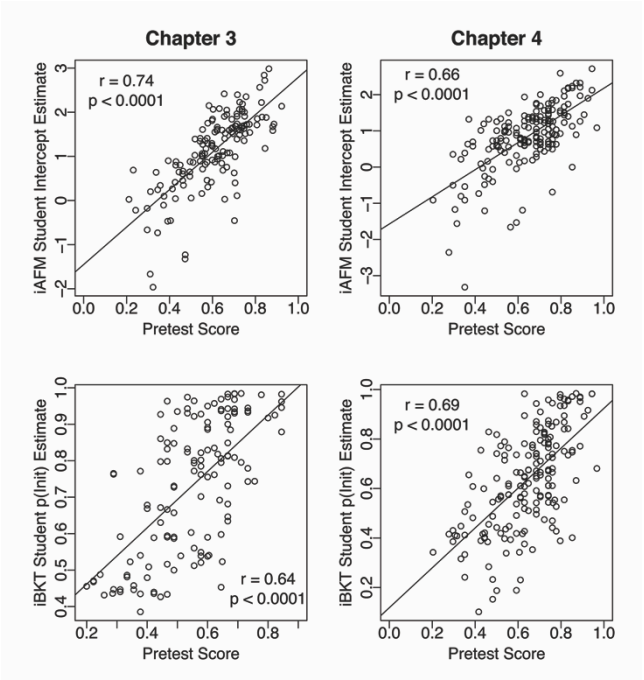

Figure 3. Relationships between out-of-tutor pretest scores and iAFM/iBKT estimates of student ability based on within-tutor data.

4.3.2 Estimates of Student Learning Rate

Given that the only external assessment data collected were a pretest and posttest, we sought to validate the construct of student learning rate (as estimated by the models) on pretest-posttest gains. Students were given roughly the same amount of time to engage with the tutors, so those with accelerated learning rates might be expected to gain more knowledge in the time available.

Thus, we examined the degree to which student learning rate estimates predicted pretest-posttest gains while controlling for pretest scores. We controlled for pretest scores because they have been shown to negatively predict learning gains due to assessment

ceiling effects. That is, students who start out performing well on the pretest have less "room for improvement".

For the Chapter 3 dataset, iAFM Student Slope (δ_i) estimates did not significantly predict learning gains. In a linear regression predicting pretest-posttest gains, pretest scores were a significant predictor (β=-0.189, p=0.005) and Student Slope estimates were not (β=0.396, p=0.144). iBKT Student p(Learn) estimates did not significant predict learning gains. In a linear regression predicting pretest-posttest gains, pretest scores were a significant predictor (β=-0.226, p=0.005) and Student Slope estimates were not (β=0.062, p=0.218).

For the Chapter 4 dataset, iAFM Student Slope (δ_i) estimates significantly predict learning gains. In a linear regression predicting pretest-posttest gains, pretest scores (β=-0.641, p<0.0001) and Student Slope estimates (β=0.576, p=0.007) were both significant predictors. iBKT Student p(Learn) estimates also significantly predict learning gains. In a linear regression predicting pretest-posttest gains, pretest scores (β=-0.645, p<0.0001) and p(Learn) estimates (β=0.133, p=0.004) were both significant predictors.

For one of the two units (Chapter 4), we observed that student learning rate estimates were validated on external assessments of learning gain. Interestingly, this is the same unit for which we observed a strong cross-model reliability in student learning rate estimates. Thus, we have converging evidence that student learning rates estimates for the Chapter 4 dataset are both reliable and valid.

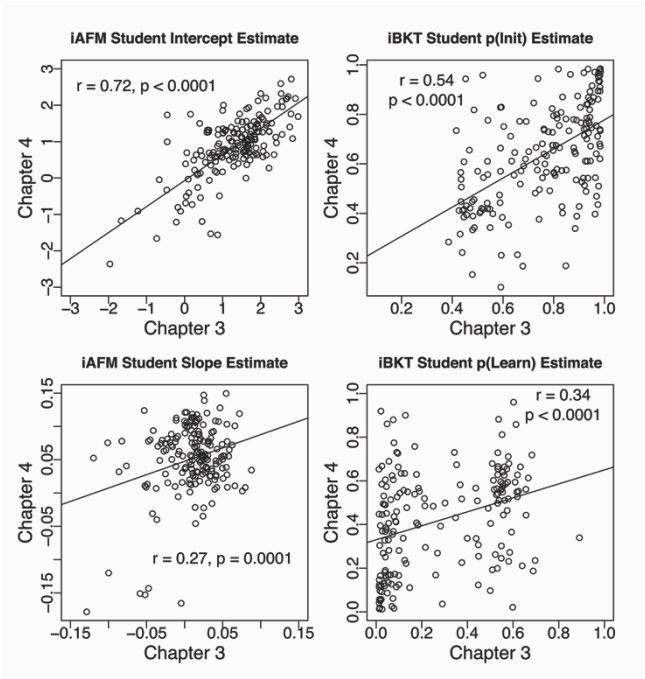

Figure 4. Relationships between student parameter estimates across the two datasets (same student population).

4.4 Towards Understanding & Using Student Parameter Estimates

4.4.1 Consistency of individual student constructs across datasets

A core motivating question for collecting two datasets on the same group of students was: How consistent are iAFM and iBKT model estimates of the student ability and student learning rate constructs across units?

Figure 4 summarizes this relationship. Estimates of student ability are fairly consistent, especially as estimated by iAFM. It seems sensible to interpret this as suggesting that overall student ability on Chapter 3 content is strongly related to overall student ability on Chapter 4 content, as we have shown estimates of student ability to be both reliable and valid.

Estimates of student learning rate are less consistent. This may either be due to the fact that Chapter 3 estimates of student learning rate were neither very reliable nor very valid. Alternatively, the differences in student learning rate estimates across the two chapters may also be due to the fact that students genuinely learn different material at different rates. Unfortunately, we cannot resolve this question with the present data. We are currently collecting more datasets from this same group of students. If we obtain more reliable and valid student learning rate estimates in future data from this group of students, we can more confidently address this question in future research.

4.4.2 Understanding student learning rate estimates

Given that we established the reliability and validity of iAFM and iBKT's parameter estimates for the Chapter 4 dataset were reasonably reliable and valid, we sought to dig deeper into the explanatory power of these estimates. To this end, we conducted exploratory analyses on the Chapter 4 data to (1) visualize the learning trajectories of students with the highest vs. lowest estimated learning rates, (2) understand the relationships between estimated learning rates and prior-knowledge and motivational factors, and (3) understand the degree of variability in estimated learning rate across students.

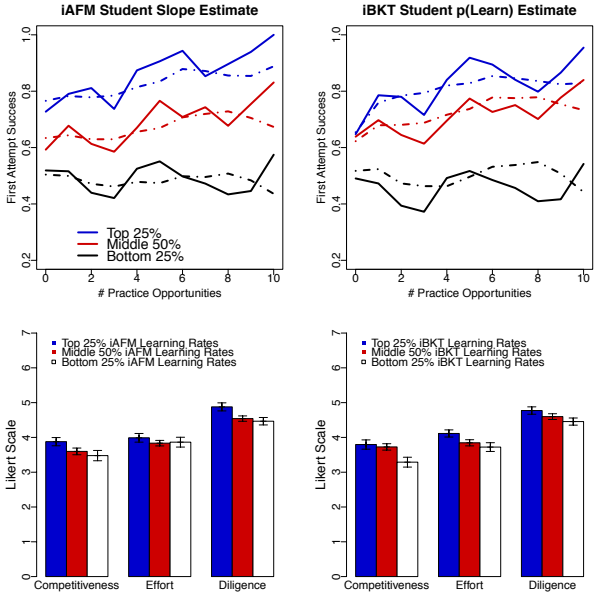

Figure 5. *Top Row*: Early-opportunity learning trajectories of students, grouped based on iAFM (Left) and iBKT (Right) estimated learning rates. Solid lines are actual data; dotted lines are each respective model's predicted performance. *Bottom Row*: Mean self-report Likert scale ratings of questions measuring dimensions of competitiveness, effort, and diligence. Grouped based on iAFM (Left) or iBKT (Right) estimated learning rates. Error bars show standard errors on the means.

Figure 5 (top row) shows the aggregate learning trajectories for students split based either on their iAFM Student Slope estimates (top left) or their iBKT Student p(Learn) estimates (top right). The top 25% of student parameter estimates are plotted in blue, the middle 50% (between 1st and 3rd quartiles) are plotted in red, and the lower 25% are plotted in black. Dotted lines represent each respective model's *predicted* earning trajectories.

One striking pattern, especially in the iAFM learning trajectories (top left), is the apparent relationship between average success on initial practice opportunities (i.e., prior knowledge) and estimated learning rate through the remaining opportunities. This observation is corroborated by a strong and significant correlation between iAFM Student Intercepts and iAFM Student Slopes ($r=0.78$, $p<0.0001$). One might interpret this to suggest that students who enter into the tutor with greater prior knowledge will be poised to gain more from the tutor (i.e., "the rich get richer"). Alternatively, students may have higher overall knowledge *because* they are fast learners. There may also be individual trait-based variables that positively drive both learning rate and overall achievement.

To explore the relationships between measures of traits relevant to learning, we analyzed self-report survey data grouped by three factors (as described in Section 3.1): Competitiveness, Effort, and Diligence. The relationship between these measures and the high, medium, and low learning rate estimates from iAFM and iBKT are shown in Figure 5 (bottom row). There appears to be a relationship between the means of each self-report measure and the general range that the learning rate estimate falls in.

We analyzed the continuous relationship between students' mean self-report rating along each dimension and their iAFM learning rate estimates. In a linear regression predicting iAFM Student Slopes, Competitiveness and Effort were not significant predictors but Diligence ($\beta=0.016$, $p=0.007$) was. In a similar linear regression predicting iAFM Student Intercepts, again Diligence was the only significant predictor ($\beta=0.02$, $p=0.04$). Thus, among self-reported measures, the strongest dimension predicting both student ability/prior knowledge *and* student learning rate was the Diligence measure. Future work using causal modeling is warranted to discover the true nature of causality among these student-level constructs.

Finally, we investigated the degree of variability in estimated learning rate across students. The first quantile of student learning rates from iAFM is 0.03 logits and the third quantile of rates from iAFM is 0.08 logits. These can be conceptualized as canonical "slow" and "fast" learners. If we were to assume starting at around 70% performance (which comes from the model's global intercept estimate), it would take the "slow" (0.03 logits) student approximately 25 opportunities to reach mastery (defined as 85%, the performance equivalent of a p(Know)=0.95, factoring in the guess and slip probabilities we used in the actual tutor). It would take the "fast" (0.08 logits) student approximately 11 opportunities to reach the same place.

4.4.3 Identifying wheel spinners

The current definition of "wheel spinning" put forth in the Educational Data Mining community is the "phenomenon in which a student has spent a considerable amount of time practicing a skill, yet displays little or no progress towards mastery" [5]. There has been some controversy around the ideal way to measure mastery (e.g., 3 corrects in a row vs. reaching a certain p(Know) in knowledge tracing). Furthermore, some students may be classified as wheel spinners based on not mastering in a certain number of opportunities but they may still be making progress.

We propose that reliable and validated estimates of individual student learning rate parameters, combined with KC learning rate parameters, could be used to estimate wheel spinning student/KC pairs in way that is agnostic to mastery status. Specifically, if the combined student and KC learning rate parameters in iAFM predict *no* improvement or negative improvement across additional practice opportunities, and aren't already at a high level of performance on their first opportunity (here we considered this to be 80% or above), we could consider the student to be wheel spinning on the KC. This method of estimating wheel spinning would be particularly useful for datasets with sparse data on some student-KC pairs, as it is not performance-dependent after the model has been fit to the full dataset.

Based on this operationalized definition, we found that approximately 15% of student-KC pairs in the Chapter 4 dataset are estimated to be wheel spinning. That is, those students are not making progress on those KCs. This is a substantially lower estimate than the 25% reported by a recent wheel spinning detector in [5]. An interesting route for future work would be to do a direct comparison of the wheel spinning detector presented in [5] and our proposed student/KC learning rate identifier within the same dataset. This would allow for testing the possibility that some students who are still making progress, albeit extremely slowly, may be prematurely labeled as "wheel spinners" by [5].

5. SUMMARY & LIMITATIONS

Previous efforts towards more explanatory, interpretable, and actionable modeling advancements in the realm of skill/knowledge component model discovery have been promising in their potential and demonstrated impact on learning science and education. The present paper represents a novel effort to bring these deeper modeling approaches, focused on ensuring explanatory power, to the realm of individualized student-parameter models.

Towards improving the reliability and validity of individualized student estimates, we collected two datasets from the same student population. Both datasets were "deep" along the dimension of student-KC observations. We fit iAFM and iBKT to both datasets and showed that the models outranked their non-individualized counterparts in terms of fit to data and predictive accuracy. Importantly, we moved beyond these metrics to show that estimates of student ability were highly reliable (iAFM and iBKT yielded strongly correlated estimates) and valid (estimates significantly predicted pretest data).

This demonstration of confidence in the student ability estimates from iBKT, but even more so iAFM, has promising implications for the possibility of individualizing the student models that determine mastery in intelligent tutoring systems at *least* in terms of overall student ability/knowledge. Our results also suggest that it would be reasonable to fix such student ability parameters, or set priors on them, based on either well-mapped pretest assessment data or prior (deep) data from those students' learning.

We also showed that estimates of student learning rate per practice opportunity were reliable and valid in one of the two datasets (Chapter 4). This is the first evidence, to our knowledge, of obtaining both reliable and valid student learning rates through a statistical model with *individualized* student parameters. We believe that this success is largely related to the amount and quality of per-student data we collected.

With the confidence of having reliable and valid parameter estimates, we then proceeded to further investigate potential explanations for differences in student learning rates within the

Chapter 4 dataset. We found a strong and significant relationship between student ability and improvement rate as well as an additional effect of diligence, based on self-report measures. Further research is warranted to distill the causal relationships between these constructs.

Knowing that a model's estimates of individualized student parameters not only fit data well, but are reliable and valid, provides greater confidence for applying the model to (1) interpret the parameter estimates to understand characteristics of students, and (2) use the model to individualize the trajectory of mastery estimation for future students.

Even though both iBKT and iAFM outperformed their non-individualized counterparts in predicting performance in the Chapter 3 dataset, we did not find strong evidence of reliability and validity of the student-specific parameter estimates. Thus, we did not rely on that dataset to help us understand individual differences in learning rates. For the same reason, we could not confidently attribute the differences, in estimated student learning rates across the datasets, to *true* differences in students' learning rates for the two chapters' material.

Although considering reliability and validity of models' parameter estimates sets a higher bar than predictive accuracy for evaluating modeling advances, we believe those to be important characteristics of a model that is to be explanatory, interpretable, and/or actionable. Here, we have demonstrated that with a sufficiently good dataset, iAFM and iBKT are individualized student models that *can* produce reliable and valid parameter estimates.

Since our present work was limited to two datasets on one population of students, it is unclear the degree to which our modeling results will generalize, especially given that at least iAFM does not produce reliable, valid parameter estimates on more sparse datasets [9]. In addition, these results are limited to two specific statistical models produce individualized estimates student-level parameters, with a particular focus on individual differences in learning rate. There are other classes of models that could be extended to estimate differences in learning rate: for example, producing individualized estimates of the differential effects of success versus failure [15]. This would be an interesting focus for future work on this topic.

Nevertheless, we have laid a foundation of methodology by which reliability and validity of parameter estimates, whether student- or KC-level, can be assessed. We have also demonstrated ways of using the reliable and valid student parameter estimates from iAFM and iBKT to yield interesting insights about student learning.

6. ACKNOWLEDGMENTS

We thank the Institute of Education Sciences for support to RL (training grant #R305B110003) and the National Science Foundation for support to Carnegie Mellon University's LearnLab (#SBE-0836012).

7. REFERENCES

[1] Aleven, V., Sewall, J., McLaren, B.M., and Koedinger, K.R. (2006). Rapid authoring of intelligent tutors for real-world and experimental use. In *Proceedings of the 6th ICALT*. IEEE, Los Alamitos, CA, pp. 847-851.

[2] Cen, H., Koedinger, K.R., & Junker, B. (2006). Learning Factors Analysis: A general method for cognitive model evaluation and improvement. Intelligent Tutoring Systems, 164-175.

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

[4] Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation, 10(7), 1895–1923.

[5] Gong, Y. & Beck, J. (2015). Towards Detecting Wheel-Spinning: Future Failure in Mastery Learning. In Proceedings of Learning At Scale '15.

[6] Koedinger, K.R., Corbett, A.C., & Perfetti, C. (2012). The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning. Cognitive Science, 36(5), 757-798.

[7] Koedinger, K.R., McLaughlin, E.A., & Stamper, J.C. (2012). Automated Student Model Improvement. 5th International Conference on EDM.

[8] Koedinger, K. R., Stamper, J. C., McLaughlin, E. A., & Nixon, T. (2013). Using data-driven discovery of better cognitive models to improve student learning. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), Proceedings of the 16th International Conference on Artificial Intelligence in Education (AIED '13), 9–13 July 2013, Memphis, TN, USA (pp. 421–430). Springer.

[9] Liu, R., & Koedinger, K. R. (2015). Variations in learning rate: Student classification based on systematic residual error patterns across practice opportunities. In O. C. Santos, J. G. Boticario, C. Romero, M. Pechenizkiy, A. Merceron, P. Mitros, J. M. Luna, C. Mihaescu, P. Moreno, A. Hershkovitz, S. Ventura, & M. Desmarais (Eds.), Proceedings of the 8th International Conference on Education Data Mining (EDM2015), 26–29 June 2015, Madrid, Spain (pp. 420–423). International Educational Data Mining Society.

[10] Liu, R., & Koedinger, K. R. (under review). Closing the loop: Automated data-driven skill model discoveries lead to improved instruction and learning gains.

[11] Liu, R., Koedinger, K. R., & McLaughlin, E. A. (2014). Interpreting model discovery and testing generalization to a new dataset. In J. Stamper, Z. Pardos, M. Mavrikis, & B. M. McLaren (Eds.), Proceedings of the 7th International Conference on Educational Data Mining (EDM2014), 4–7 July, London, UK (pp. 107–113). International Educational Data Mining Society.

[12] Lee, J.I., & Brunskill, E. (2012). The Impact on Individualizing Student Models on Necessary Practice Opportunities. 5th International Conference on EDM.

[13] Pardos, Z.A., & Heffernan, N.T. (2010). Modeling individualization in a bayesian networks implementation of knowledge tracing. User Modeling, Adaptation, and Personalization, 255-266.

[14] Pardos, Z. A., Trivedi, S., Heffernan, N. T., & Sárközy, G. N. (2012). Clustered knowledge tracing. In S. A. Cerri, W. J. Clancey, G. Papadourakis, K.-K. Panourgia (Eds.), Proceedings of the 11th International Conference on Intelligent Tutoring Systems (ITS 2012), 14–18 June 2012, Chania, Greece (pp. 405–410). Springer.

[15] Pavlik, P.I., Cen, H., & Koedinger, K.R. (2009). Performance factors analysis–a new alternative to knowledge tracing. AIED, 531–538.

[16] Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. doi:10.1214/10-STS330

[17] Stamper, J., & Koedinger, K. R. (2011). Human-machine student model discovery and improvement using data. *Proceedings of the 15th International Conference on*

Artificial Intelligence in Education (AIED '11), 28 June–2 July, Auckland, New Zealand (pp. 353–360). Springer.

[18] Yudelson, M.V., Koedinger, K.R., & Gordon, G.J. (2013). Individualized bayesian knowledge tracing models. AIED, 171-180.

[19] VanLehn, K. (2006). The behavior of tutoring systems. *International Journal of Artificial Intelligence in Education, 16*, 227–265.