Variations in learning rate: Student classification based on systematic residual error patterns across practice opportunities

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

A growing body of research suggests that accounting for student-specific variability in educational data can improve modeling accuracy and may have implications for individualizing instruction. The Additive Factors Model (AFM), a logistic regression model used to fit educational data and discover/refine skill models of learning, contains a parameter that individualizes for overall student ability but not for student learning rate. Here, we show that adding a per-student learning rate parameter to AFM overall does not improve predictive accuracy. In contrast, classifying students into three "learning rate" groups using residual error patterns, and adding a per-group learning rate parameter to AFM, substantially and consistently improves predictive accuracy across 8 datasets spanning the domains of Geometry, Algebra, English grammar, and Statistics. In a subset of datasets for which there are pre- and post-test data, we observe a systematic relationship between learning rate group and pre-to-post-test gains. This suggests there is both predictive power and external validity in modeling these distinct learning rate groups.

Keywords

Student learning rate, learning curves, Additive Factors Model

1. INTRODUCTION

A growing body of research suggests that accounting for student-specific variability in statistical models of educational data can yield prediction improvements and may potentially inform instruction. The majority of work investigating the effects of student-specific parameters [6, 10, 11, 15] has been done in the context of a class of models called Bayesian Knowledge Tracing (BKT), a special case of using Hidden Markov Models to model student knowledge as a latent variable.

Logistic regression is another popular method for modeling educational data. The Additive Factors Model (AFM) [4] is one instantiation of logistic regression that was developed with the primary intention of evaluating, discovering, and refining *knowledge component (KC) models* (also referred to as Q-matrices). In contrast to *statistical models* of educational data, KC models define the knowledge components (e.g., skills, concepts, facts) on which estimates of students' knowledge are based. AFM has parameters modeling KC difficulty, KC learning rate, and individual student ability, but it does not have a parameter for individual student *learning rate*.

Recent work extending BKT models [15] suggests that better predictive accuracy is achieved by adding parameters that accommodate different learning rates for different students. Here, we investigate two different extensions of AFM that model student learning rate variability. The first model (AFM+StudRate) adds a per-student learning rate parameter to AFM, dramatically increasing the number of parameters in the model. We find some evidence that this model overfits the training data. For the second model (AFM+GroupRate), we introduce a method of classifying students into learning rate groups. We then add a per-group, rather than per-student, learning rate parameter to AFM and show that this model significantly outperforms regular AFM in predictive accuracy across 8 datasets spanning various domains.

Importantly, we move beyond simply evaluating the models in terms of their predictive accuracy to assess the external validity of the additional parameters. We show that they relate significantly to post-test outcomes. Validation and interpretation of statistical model parameter fits are a critical step towards successfully bridging EDM, the science of learning, and instruction.

1.1 The Additive Factors Model

AFM is a logistic regression model that extends item response theory by incorporating a growth or learning term. 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 difficulty (β_k) of the required knowledge components or KCs on that problem step (Q_{jk}), and the improvement (γ_k) in each of the required KCs with each additional practice opportunity multiplied by the number of practice opportunities (T_{ik}) the student has had with that KC prior to the current problem step [4].

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

AFM accommodates some individualization with the student ability parameter but makes the simplifying assumption that students learn at the same *rate*, since the original purpose of AFM was to refine KC models [4]. Here, we investigate whether extensions of AFM can accommodate variability in student learning rates and provide meaningful information about learning rate differences.

2. IDENTIFYING AND MODELING LEARNING RATE VARIATION

To explore adding learning rate variation to AFM, we created two new models extending AFM. The first model (AFM+StudRate) adds a per-student learning rate parameter, and the second model (AFM+GroupRate) adds a per-group learning rate parameter whereby membership among the three groups is determined using the method described in Section 2.1.

2.1 Student classification method

To classify students, we sought to identify those who improve—with each practice opportunity—more (or less) so than would be predicted by traditional AFM, which has a per-KC rate parameter that already accounts for the learning rate variability that is predicted by the KCs present at each opportunity. To do so, we examined the patterns in residual errors across opportunity counts after the data are fit with traditional AFM. A student whose learning curve is steeper than that predicted by AFM will exhibit systematically increasing residual errors; i.e., residuals will correlate positively with opportunity count. Conversely, a student whose performance consistently increases *less* per opportunity than AFM predicts will exhibit a negative correlation between residual error and opportunity count.

To leverage this feature of residual error to classify students, we first fit the baseline AFM model to a full dataset (all students and KCs). Then, for each individual student, deviance residuals were computed, comparing the AFM model prediction against the actual data. Correlation coefficient cut-offs were set for each dataset at r > 0.1 for the "steep" learning-curve group and r < -0.1 for the "flat/declining" learning-curve group. Based on exploratory analyses, we selected the most stringent cut-off that yielded reasonable group sizes (approximately 50% students classified into either the steep or flat groups). The remaining students, whose learning curves were reasonably captured by the per-KC learning rates specified in AFM, were classified into a third "regular" group.

2.2 AFM+StudRate and AFM+GroupRate

The model that extends AFM by adding a per-student learning rate (AFM+StudRate) is given in Equation 2. It contains the parameters of traditional AFM with an additional parameter capturing the improvement (δ_i) by each student with every additional practice opportunity. Here, T_{ik} represents the practice opportunity count of a given KC required for a problem step *j*.

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

The model that extends AFM by adding a per-group learning rate (AFM+GroupRate) is given in Equation 3.

$$\ln\left(\frac{p_{ij}}{1-p_{ij}}\right) = \theta_i + \sum_{k\in KCs} Q_{jk}(\beta_k + \gamma_k T_{ik} + \delta_c S_{ic} T_{ik}) \qquad (3)$$

It uses the same parameters as AFM+StudRate except that each student's improvement rate with each additional practice opportunity (δ_c) is derived from a per-group rate (and thus can only take on one of three different values). Each student's group membership is specified by S_{ic}, which takes on a value of 1 when the student *i* belongs to group *c* and a value of 0 otherwise.

3. EVALUATING MODELS FOR FIT AND PREDICTIVE ACCURACY

3.1 Datasets

To test these statistical models on real educational data and to compare their predictive accuracies, we applied them across 8 datasets from DataShop [8]: Geometry Area 96-97, Cog Model Discovery Experiment Spring 2010, Cog Model Discovery Experiment Spring 2011, Cog Model Discovery Experiment Fall 2011, Assistments Math 2008-2009 Symb-DFA, Self Explanation sch_a3329ee9 Winter 2008 CL, IWT Self-Explanation Study 1 Spring 2009, and Statistical Reasoning and Practice - Fall 2009. These span a variety of content domains: Geometry, Equation solving, Story problems, English grammar, and Statistics. All of these datasets are publicly available at http://pslcdatashop.org.

We selected datasets that had already undergone significant KC model refinement via both manual and automated methods [9].

3.2 Methods

Each dataset was pre-processed based on the single-skilled KC model that achieved the best item-stratified CV performance according to values reported on DataShop. Table 1 lists the names of the KC models used and the number of KCs in each model. The three AFM models were implemented in R with student ability (θ_i), KC difficulty (β_k), and all learning rate parameters modeled as random effects, since many datasets used here were characterized by non-uniform sparsity in student-KC pairings, due to the mastery-based adaptive nature of the tutors from which the data originate. Modeling the parameters as random effects also reduces the likelihood of over-fitting the data by keeping their estimates close to zero.

The sparsity found in mastery-based datasets is particularly extreme at high opportunity counts, and this introduces noise to our classification method, which is dependent on good resolution across opportunity counts. Thus, we employed a conservative and systematic opportunity count cut-off method prior to analyses. The number of observations at each opportunity count was totaled for each student. Counts at which the average observations per student was less than 1 *and* the number of observations for any single student was 1 or fewer were excluded. In other words, at the excluded opportunity counts, no student had more than 1 *total* observation, and the majority of students did not have any. This excluded a very small percentage of total observations; the percent of observations retained are reported in the "Opp Cut-off" column of Table 1. In addition, our grouping technique required at least 5 observations in order to run the residual-by-opportunity correlations, so students who performed fewer than 5 total problem steps were excluded from the analyses. The left-most column of Table 1 reports the number of students included (with the original N in parentheses).

Models were evaluated for each dataset using Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and cross-validation measures. Two types of cross-validation (CV) were assessed: item-stratified CV, in which different random folds contain different problem steps, and student-stratified CV, in which different random folds contain different students (i.e., the model is tested on "unseen" students). Due to the random nature of the folding process, we repeated ten runs of each type of 10-fold CV, and the mean RMSEs across each run were used to compute the overall means and standard errors (in parentheses) reported in Table 2. Any CV results in which AFM+StudRate or AFM+GroupRate significantly outperforms regular AFM (as assessed by p<0.05 in a paired t-test between mean RMSEs across the 10 runs) are denoted with stars.

3.3 Results

The results of fitting the three statistical models to all 8 datasets are summarized in the right-most columns of Table 1.

AFM with a per-student learning rate fails to perform consistently better than regular AFM either across metrics within any dataset or across datasets. With an extra parameter per student, AFM+StudRate naturally fits training data better, but the evaluation metrics indicate over-fitting that is likely idiosyncratic (i.e., resulting in parameter estimates that will not generalize well to "unseen" items or students). Even for the AIC metric, which incorporates a smaller penalty for extra parameters than BIC, AFM+StudRate is better than regular AFM for only half of the datasets and only slightly so. By BIC, it is better than regular AFM in only one dataset. Cross-validation reveals that AFM+StudRate fails to achieve significantly lower RMSEs than regular AFM in 14 of 16 cases.

In contrast, AFM+GroupRate performs best on *all* 8 datasets by AIC, BIC, and item-stratified CV measures. It also performs the best on the majority of datasets (6 out of 8) by student-stratified CV. The superior performance according to student-stratified CV is particularly notable, because the predictions are made on data from "unseen" students. That is, no student information (not even group membership) is available for the data in the test set. The fact that AFM+GroupRate performs better than regular AFM implies that this model is successfully capturing some student-level variability that produces better, cleaner KC parameters. This is not true for AFM+StudRate, which did not achieve significantly better student-stratified CV for any dataset.

4. RELATIONSHIP TO PRE-POST GAINS

Predictive accuracy is often used as a proxy for quality in EDM models. Assessing the validity of these student groups beyond relevance to model-fitting is equally, if not more, important. To do so, we investigated the relationship between group membership and post-test outcomes. Four of the datasets tested in Section 3 contained pre/post-test data that were accessible via DataShop: the three geometry Cog Discovery datasets and the IWT 1 dataset.

For each dataset we ran a simple regression with both pre-test score and per-group coefficients (from fitting AFM+GroupRate) as predictors of post-test score. Even after taking into account the variance explained by pre-test scores, learning rate group membership predicts post-test scores significantly for Cog Discovery Spring 2010 ($p<0.001$), Cog Discovery Fall 2011 ($p=0.016$), and Cog Discovery Spring 2011 ($p<0.001$), and marginally significantly for IWT 1 ($p=0.077$). These results suggest that group classification predicts unique variance in post-test outcomes and is thus a valid and interpretable construct.

5. DISCUSSION

5.1 Conclusions and implications

In the present work, we investigated two extensions of AFM that incorporated learning rate variation: adding a per-student learning rate parameter (AFM+StudRate) and adding a per-group learning rate parameter (AFM+GroupRate). AFM+StudRate overall did not significantly improve upon regular AFM, according to predictive accuracy metrics. In contrast, the residual-based student grouping method we developed seems to capture meaningful differences in learning rate variations. The groups have internal validity: adding a per-group learning rate to AFM improved predictive accuracy across all datasets based on the vast majority of fit metrics. They also have external validity: per-group rate

Table 1. *Dataset details and predictive accuracy metrics for each of the three statistical models fit to datasets. The percent of observations retained for analyses are shown in parentheses underneath opportunity cut-off values. Item- and student-stratified CV values are mean RMSEs over 10 separate runs of 10-fold cross validation, with standard errors in parentheses. Stars denote models with significantly better cross-validation performance (at p<0.05 in paired t-tests of RMSE values across CV runs) than regular AFM. The best-performing models by each metric are bolded.*

Dataset [Domain] # Students	KC Model (# KCs)	Opp Cut-off	Statistical Model	AIC	BIC	Item-Strat CV RMSE	Student-Strat CV RMSE
Geometry **1996-97** [Geometry] N = 56 (of 59)	LFASearchAIC WholeModel3 (18)	27 (99.22%)	AFM	5039.7	5072.4	.3996 (.0003)	**.4063 (.001)**
			+StudRate	5043.8	5080.5	.3991 (.0004)	.4063 (.001)
			+GroupRate	**4999.2**	**5038.4**	**.3975 (.0003)***	.4068 (.001)
Cog Discovery **Spring 2010** [Geometry] N = 123 (of 123)	KTskills.Mcontext.s ingle.sep.ind.areas (42)	80 (99.72%)	AFM	29208.5	29251.7	.3238 (.00003)	.3319 (.0001)
			+StudRate	29160.8	29221.3	.3232 (.00002)*	.3318 (.0001)
			+GroupRate	**29030.1**	**29081.9**	**.3230 (.00002)***	**.3317 (.0001)***
Cog Discovery **Spring 2011** [Geometry] N = 65 (of 69)	KTracedSkills.matc hed.Fall2011 (7)	30 (99.3%)	AFM	4099.7	4131.5	.3877 (.0002)	.4025 (.0004)
			+StudRate	4101.4	4146.0	.3879 (.0002)	.4025 (.0004)
			+GroupRate	**4077.3**	**4115.3**	**.3856 (.0002)***	**.4017 (.0004)***
Cog Discovery **Fall 2011** [Geometry] N = 103 (of 103)	KTracedSkills.Conc atenated (15)	26 (97.87%)	AFM	3175.9	3208.2	.3104 (.0003)	**.3194 (.0003)**
			+StudRate	3177.8	3223.0	.3108 (.0003)	.3198 (.0003)
			+GroupRate	**3155.6**	**3194.3**	**.3090 (.0002)***	.3198 (.0003)
Assistments **Symb-DFA** [Story Problems] N = 318 (of 318)	Main.LFASearch Model0 (4)	11 (98.81%)	AFM	6013.1	6046.0	.4265 (.0006)	.47008 (.0001)
			+StudRate	6016.9	6062.9	.4267 (.0006)	.47008 (.0001)
			+GroupRate	**5793.2**	**5832.7**	**.4166 (.0006)***	**.47005 (.0001)**
Self-Explanation **Winter 2008** [Equation Solving] N = 70 (of 71)	LFASearchAIC Model.r2 (19)	49 (98.78%)	AFM	6201.8	6235.6	.3905 (.0002)	.4140 (.0005)
			+StudRate	6201.4	6248.8	.3906 (.0002)	.4141 (.0006)
			+GroupRate	**6158.9**	**6199.5**	**.3889 (.0002)***	**.4127 (.0005)***
IWT 1 **Spring 2009** [English Grammar] N = 120 (of 120)	LFASearchAIC WholeModel1 (26)	11 (98.64%)	AFM	6820.8	6854.7	.4134 (.0003)	.4392 (.0002)
			+StudRate	6815.2	6862.7	.4128 (.0003)*	.4392 (.0002)
			+GroupRate	**6752.9**	**6793.6**	**.4099 (.0002)***	**.4389 (.0002)***
Statistics **Fall 2009** [Statistics] N = 52 (of 52)	LFASearchAIC Model0 (16)	30 (99.81%)	AFM	2967.8	2999.4	.3090 (.0032)	.3250 (.0003)
			+StudRate	2965.5	3009.8	.3105 (.0031)	.3250 (.0004)
			+GroupRate	**2935.5**	**2973.5**	**.3085 (.0029)***	**.3248 (.0003)***

coefficients significantly predict each group's post-test outcomes, controlling for pre-test.

Despite the focus of the AFM+GroupRate model on student-level differences, adding the per-group rate parameter produces more accurate estimates of KC parameters, based on the model's superior performance in student-stratified CV for the vast majority of datasets. The only information the model gets for fitting test data in student-stratified CV ("unseen" students whom the model has no information about with respect to ability, learning rate, or group) are the KC parameters. For this reason, AFM+GroupRate may be useful for data-driven refinement of KC parameters, which in turn has implications for instruction (e.g., parameter-setting in Knowledge Tracing based cognitive tutors [14]).

Compared to other statistical models extending AFM (Performance Factors Analysis [12], Instructional Factors Analysis [5], Recent Performance Factors Analysis [7]), AFM+GroupRate adds relatively few parameters (only three) to AFM but achieves consistent and substantive improvements in prediction. These three parameters' coefficient estimates are consistently interpretable (the per-group learning rates are ordered according to intuitions about each group's learning curve steepness), and the model avoids overloading on the interpretation of parameters.

We conducted extensive post-hoc analyses to interpret what the three learning groups actually reveal about student behavior and did not find evidence that the groups detect learning speed as an inherent trait, per se. For example, high ability students did not tend to be in the "steep" group, and low ability students did not tend to be in the "flat" group. Rather, the amount of improvement per opportunity seems to differ, more generally, depending on where the learner is on his/her *true* learning curve for any given skill. That is, the improvement per opportunity may be different for the earliest opportunities on a skill than for much later opportunities on a skill. Different students' learning curves within cognitive tutor data may vary because they start using the cognitive tutor at different points of their true learning curves for any given skill, depending on their experience with that skill prior to tutor use. We found evidence supporting this notion in post-hoc analyses. Considered in conjunction with the lack of evidence for a per-student learning rate, our findings contradict the intuitive notion that some students naturally learn faster than others.

5.2 Limitations and future work

The present results somewhat conflict with a finding from [15] that adding a per-student learning rate parameter to BKT yields substantial improvements in model fit, though we note that that report did not provide an interpretation nor any external validity evidence. We did not observe a benefit when adding a per-student learning rate parameter to AFM. Further work to compare these per-student parameter estimates across AFM and BKT and to externally validate the estimates from individualized BKT will provide insight into this issue.

Based on our post-hoc analyses, classification into the "flat/declining" group seems to capture high-ability students who descend into noisy performance at late opportunity counts (indicating boredom and/or "gaming the system" [2]) and low-ability students who never seem to improve ("wheel spinners" [1]). It would be interesting to validate this by seeing whether the detectors in [1] and [2] yield the same students when tested within the present datasets.

Another avenue for future investigation is to assess the degree to which different learning rate groups would benefit optimally from different KC models, via KC model search (as in [13]).

6. ACKNOWLEDGMENTS

We thank Carnegie Learning, Inc. for providing the majority of the datasets that were analyzed, Christopher MacLellan for insightful discussion, IES for support to RL (training grant #R305B110003), and NSF for support to LearnLab (#SBE-0836012).

7. REFERENCES

[1] Beck, J.E., & Gong, Y. (2013). Wheel-spinning: Students who fail to master a skill. AIED, 431-440.

[2] Baker, R.S.J.d., Corbett, A.T., Roll, I., & Koedinger, K.R. (2008). Developing a generalizable detector of when students game the system. User Modeling and User-Adapted Interaction, 18(3), 287-314.

[3] Baker, R.S.J.d., D'Mello, S.K., Rodrigo, M.M.T., & Graesser, A.C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. International Journal of Human-Computer Studies, 68(4), 223-241.

[4] 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.

[5] Chi, M., Koedinger, K.R., Gordon, G., Jordan, P., & VanLehn, K. (2011). Instructional factors analysis: A cognitive model for multiple instructional interventions. 4th International Conference on EDM.

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

[7] Galyardt, A., & Goldin, I. M. (accepted). Move your lamp post: Recent data reflects learner knowledge better than older data. Journal of Educational Data Mining.

[8] Koedinger, K.R., Baker, R.S., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A Data Repository for the EDM community: The PSLC DataShop. In Romero, C., Ventura, S., Pechenizkiy, M., Baker, R.S.J.d. (Eds.) Handbook of Educational Data Mining. Boca Raton, FL: CRC Press.

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

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

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

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

[13] Rafferty, A.N., & Yudelson, M. (2007). Applying learning factors analysis to build stereotypic student models. Frontiers in Artificial Intelligence and Applications, 158, 697.

[14] Ritter, S., Anderson, J.R., Koedinger, K.R., & Corbett, A. (2007). The Cognitive Tutor: Applied research in mathematics education. Psychonomics Bulletin & Review, 14(2), 249-255.

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