Convergent Validity of a Student Model: Recent-Performance Factors Analysis

Ilya Goldin
Center for Digital Data, Analytics, and Adaptive Learning
Pearson
ilya.goldin@pearson.com

April Galyardt
University of Georgia
110 Carlton St.
Athens, GA
galyardt@uga.edu

ABSTRACT

Models of student performance can incorporate a skill decomposition that lists the skills that each activity requires. A good model must be sensitive to improvements in skill decomposition. We validate the Recent-Performance Factors Analysis model of student performance by checking its sensitivity to the skill decomposition. We use a dataset from a tutoring system where the skill model has been improved by the Learning Factors Analysis algorithm for skill model refinement and by expert validation. We find that R-PFA reflects improvements in the skill model, providing evidence of convergent validity of R-PFA. We argue that R-PFA may be sensible as a predictive model in Learning Factors Analysis because of its convergent validity and because the R predictor of R-PFA represents mastery-aligned learning curves.

1. INTRODUCTION

Predictive models of student performance often incorporate a skill model. For example, the Additive Factors Model [3] embeds a Q-matrix [11, 1] to relate prior practice on a skill to subsequent practice on the same skill. Bayesian Knowledge Tracing [4] similarly uses a skill model in that all BKT parameters are specific to a skill.

A skill model annotates instructional activities in terms of the skills that the activities require. This tagging can be wrong, or at least suboptimal, degrading instruction in several ways. For instance, if the tagging fails to distinguish two skills, it will treat all assessments of the two separate skills as assessments of one combined skill. In fact, because a student may have differential mastery of the two skills, the combined assessment may cause a tutoring system to call for extraneous practice for one skill, and insufficient practice for another. It follows that the refinement of a skill tagging of activities can advance instruction and assessment.

When a predictive model of student performance incorporates a skill model, we can validate the performance model by seeing if it is sensitive to changes in the skill model. A learning curve represents the "power relationship between the error rate of performance and the amount of practice" [3], plotting average error across students at every practice opportunity. If the curve treats a whole curriculum as one skill, its slope will be flat, because there will be both drops and spikes in the error rates as students learn one part of the curriculum after another. If we plot separate curves for distinct skills, their slopes will not be flat, corresponding to error rates dropping as students learn. This is the intuition for the Learning Factors Analysis algorithm [3], which searches the space of possible refinements to a skill model.

Prior study of representations of recent student performance, including box and exponential kernels with a range of bandwidths, produced the Recent-Performance Factors Analysis (R-PFA) model [6, 5]. In the recency representations with the highest predictive accuracy, the weight given to the each observation decreased with the age of the observation, placing $\sim 50\%$ of weight on the last 2 attempts, and $\sim 80\%$ on the last 5. This optimal weighting was consistent across real data and a variety of simulated student behaviors.

The current work validates R-PFA by checking whether its fit to data is improved by sensible changes to the skill tagging in a dataset. The following section describes a dataset and its multiple skill models, and presents R-PFA and several comparison models. The subsequent section reports that R-PFA and the other models are all sensitive to improved skill tagging, but R-PFA has the highest predictive accuracy among the models. Finally, we discuss how R-PFA may be interpreted as representing mastery-aligned learning curves [8], and R-PFA may fit within the Learning Factors Analysis algorithm for skill model refinement.

2. METHODS

We evaluate R-PFA on a dataset in which the skill tagging has been well-studied and revised [7], originating from Cognitive Tutor Geometry by Carnegie Learning [10, 2]. This tests R-PFA in two ways; first, how will R-PFA perform in terms of predictive accuracy? Second, does R-PFA agree with prior refinement of the skill model in this dataset [7]?

This Geometry dataset has three skill models that vary in how they treat "forward" and "backward" computations of area of geometric figures [7]. The original tagging (called Merged) separates area computation by geometric shape (square, circle, etc.), but merges together forward and backward computation. The Circle-Square tagging has separate

skills for the forward and backward computations for circles and squares. The Distinct tagging has separate forward and backward skills for each of many shapes. The geometry data set contains 38,426 unique actions by 82 students. The total number of skills in each tagging is 56 in Merged, 58 in Circle-Square, and 66 in Distinct.

We compare R-PFA to baseline models Item Response Theory 1PL, Additive Factors Model [3] and Performance Factors Analysis [9] (Eqs. 1-4). All student and skill intercepts and slopes are "random", that is, drawn from a common distribution. Treating skill parameters as random "borrows strength" for their estimation by proposing that infrequently practiced skills ought to have similar parameters as skills for which more data are available. Notation: j indexes skills, i indexes students, t indexes practice opportunities. T_{ijt} is the count of prior practice, S_{ijt} is the count of prior successes, and F_{ijt} is the count of prior failures.

$$\text{IRT 1PL} \quad \theta_i + \beta_j \tag{1}$$

$$\text{AFM} \quad \theta_i + \beta_j + \gamma_j T_{ijt} \tag{2}$$

$$\text{PFA} \quad \theta_i + \beta_j + \alpha_j S_{ijt} + \rho_j F_{ijt} \tag{3}$$

$$\text{R-PFA} \quad \theta_i + \beta_j + \delta_j R_{ijt} + \rho_j F_{ijt} \tag{4}$$

R_{ijt} is the proportion of recent successes in R-PFA (Eq. 5):

$$\text{exponential kernel } R_{ijt} = \frac{\sum_{p=-2}^{t-1} d^{(t-p)} X_{ijp}}{\sum_{p=-2}^{t-1} d^{(t-p)}} \tag{5}$$

3. RESULTS AND DISCUSSION

3.1 Predictive Accuracy

We compare predictive model accuracy in terms of AIC, a metric that rewards models for predictive accuracy and penalizes them for using excessive parameters. AIC is comparable to cross-validation with a prediction error loss function, but is more appropriate for sparse datasets, such as when only a handful of students may practice a skill [6].

Table 1: Predictive accuracy (lower AIC is better).

	Skill tagging		
Model	Merged	Cir-Sq	Distinct
IRT 1PL	21652	21538	21523
AFM	21373	21252	21272
PFA	21326	21197	21211
exp R-PFA r(0.7), f(0.1)	21142	20969	21003
exp R-PFA "best" from search	21134	20949	20977
"best" decay rates: R, F	0.7, 0.3	0.5, 0.3	0.4, 0.3

For all 3 skill taggings, R-PFA has higher predictive accuracy than the other models, with PFA, AFM, and Item Response Theory 1PL following in that order (Table 1). IRT 1PL has the lowest predictive accuracy, likely because it does not reflect learning over time. At the best-performing R and F decay weights from prior work (0.7 and 0.1, respectively), the number of parameters in PFA and R-PFA is exactly the same, and R-PFA's advantage in AIC over PFA is due to increased predictive accuracy.

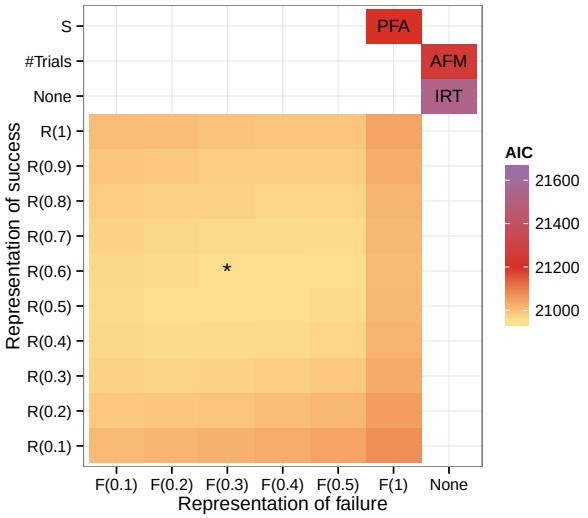

Figure 1: AIC for all 63 models on Circle-Square tagging. * denotes the best overall model.

Searching over decay rates shows that R-PFA is robust to a range of rates (Fig. 1). Even though the strictly lowest AIC uses decay rates that differ from prior work, this effect is smaller (26 points on Distinct, Table 1) than the effect of using R-PFA over other models or of improving the skill tagging, and R-PFA's performance degrades gracefully. Tuning decay rates separately for skill models has only a marginal benefit, and may confound skill model comparison.

We compare the learning curves of the 4 performance models (Fig. 2 and 3), omitting practice opportunities with fewer than 5 students. The red curves show the empirical percent correct at each opportunity, with a binomial 95% Bayesian credible interval that uses a Jeffreys prior. For example, at the 1st opportunity for circle-area backward, the mean is 45% correct, with CI (21%, 41%). The intervals make no adjustment for multiple comparisons (at each practice opportunity), so they are overly narrow, but remain useful for comparing model predictions to student performance.

The model fit curves (black) show the 2.5th and 97.5th quantiles of the model predictions. A model should predict that some students have a lower probability of a correct answer than the population percent correct, and other students, respectively, have a higher probability. If a model fits the data well, the black model curves should be centered over the empirical red curves, but should have wider bars on early attempts where there are many students in the sample.

R-PFA consistently tracks the empirical learning curve more closely than the alternative models for all 6 skills, but most clearly in circle-area backward and square-area backward (Fig. 2). Consider AFM and R-PFA predictions on circle-area backward opportunity 1: AFM predicts that 60% of students will respond correctly, when only 45% do; in fact 95% of model predictions for AFM are above the empirical percent correct. AFM produces many false positives on this early opportunity. For R-PFA, the model predictions are

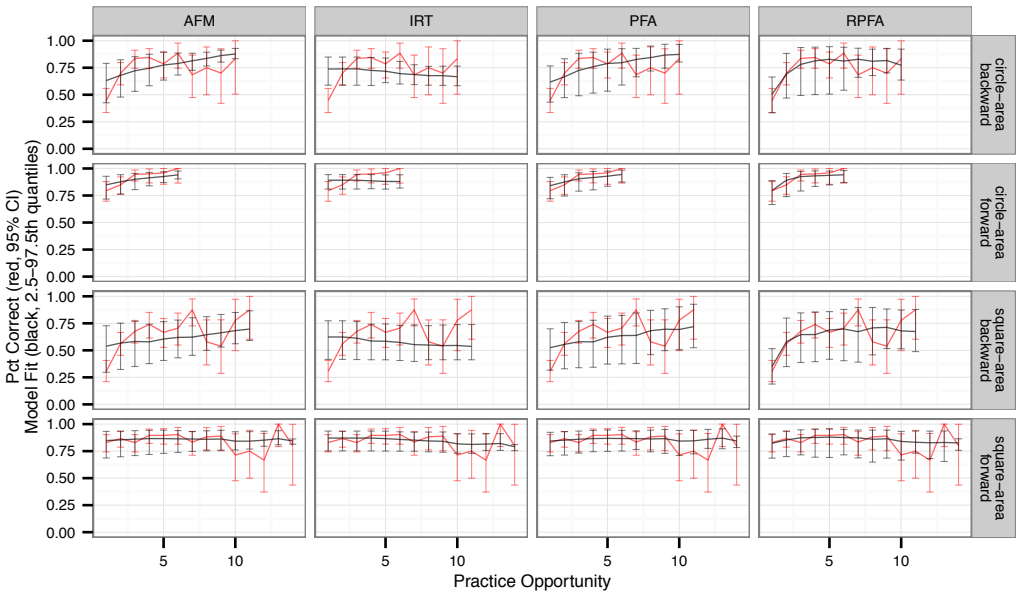

Figure 2: Empirical learning curves (red) and model fits (black) for newly split skills tagged in Cir-Sq.

centered over the empirical percent correct, and producing fewer false positives. On opportunity 4, AFM predictions are too low. AFM underestimates the amount of learning that has occurred, while R-PFA predictions track the empirical percent correct. Moreover, the R-PFA predictions range from below 0.5 to above 0.9, indicating that R-PFA is able to distinguish students who have learned the skill from those who need more practice.

3.2 Sensitivity to Skill Tagging

All models except IRT 1PL (which has the worst AIC) replicate the ranking of the three skill taggings [7]. The Cir-Sq tagging provides the best balance of predictive accuracy and data fit, compared to the Distinct tagging (which may be more granular than necessary to describe this dataset), and the Merged tagging (not sufficiently granular). While both the tagging and R-PFA are merely imperfect models, the replication provides convergent evidence for the validity of both. Skill model refinement need not improve predictive accuracy, but if it does and if the refinement makes sense in terms of instruction and cognition, that provides some evidence that the change represents an aspect of learning that is reflected in student performance.

R-PFA with the Merged tagging has a lower AIC score than any other model with the Cir-Sq tagging. Even though the Cir-Sq split is sensible and R-PFA benefits from it, R-PFA is more robust to the absence of such a split than other models. This shows in R-PFA's fit to the learning curve of circle-area (Fig. 3). AFM's predictions do not reflect the performance drop on opportunities 11 and later, but R-PFA does. This decrease motivated splitting circle-area into forward and backward skills, as in Cir-Sq [7], but R-PFA hews to the curve even without the split.

3.3 R-PFA Disaggregates Learning Curves

R-PFA effectively disaggregates the learning curves of individual students. Traditional learning curves are aligned at the first practice opportunity. Mastery-aligned curves [8] are aligned in terms of the opportunity at which students first achieve mastery. Traditional curves may conceal learning, such as if students differ in their relevant skill knowledge before their first observed practice opportunity, or if a skill model conflates two distinct skills [8]. The proportion of recent successes R by itself is a decay-weighted moving average that represents (in a non-parametric, non-model based way) the probability of mastery. R reflects the mastery-aligned curve in a predictive model, analogous to how total practice T represents the traditional learning curve in AFM.

The slope of R in R-PFA requires a different interpretation than the slope of T. A history of practice where recent success is positively associated with subsequent success (and recent failure is positively associated with subsequent failure) will have a positive slope, i.e., a positive effect on predicting the outcome. Practice relatively far in the past, whether successful or not, will have comparatively little effect on the prediction. (With the decay rate $d = 0.7$, practice older than about 5 opportunities has little effect on the prediction [6].)

One case in which the direction of the slopes of R and T may differ is in the case of a "blip" [4], i.e., when two skills follow each other in one curve, and the success rate drops in the middle of the curve, corresponding to the beginning of practice on a second skill (circle-area in Fig. 3). The slope of T ought to be flat in such a circumstance, which has been taken to mean that the skill may require a split. The slope of R will be positive, representing the fact that there is learning along the first disaggregated curve, and then along the second disaggregated curve. In fact, slopes of circle-area

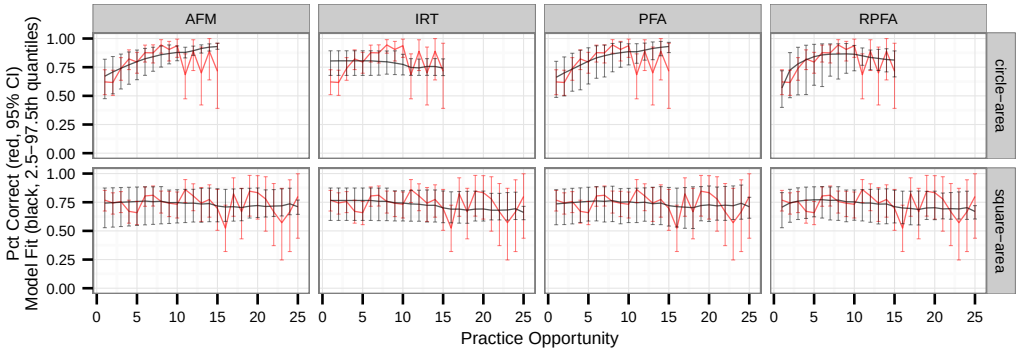

Figure 3: Learning curves (red) and model fits (black) for skills tagged in Merged, but later split in Cir-Sq.

are positive according to both AFM and R-PFA, and slopes of square-area are flat according to both AFM and R-PFA, suggesting that the slope of T is not an ideal heuristic for choosing a skill for a split.

An alternative heuristic is that when the slope of R is negative or flat, that implies that even disaggregated, mastery-aligned learning curves are a poor representation of the skill at hand. This suggests issues with the tagging of problems for this skill. This is a reasonable opportunity to invite experts to investigate "difficulty factors" for this skill, and to use LFA to apply these factors.

4. CONCLUSIONS

This investigation validates the R-PFA model of student performance in predictive accuracy on a real-world dataset. It provides convergent validity evidence for R-PFA by showing that it is sensitive to changes in a well-documented skill tagging, and yet robust to noise in a skill model. Given that no skill model is perfect, a predictive model that is accurate even in the face of such noise could be an asset to adaptive learning technologies.

The skill tagging refinement algorithm LFA [3], which incorporates AFM, may benefit by switching to R-PFA. LFA uses AFM in two ways: as a component in A* search, and as an interpretable learning curve slope. R-PFA may be a better component in A* search, because it is a more accurate model that is still sensitive to skill model changes, and because it reflects a mastery-aligned curve rather than an aggregate curve. The interpretation of the slope parameter is different, but sensible.

5. ACKNOWLEDGMENTS

We thank Ran Liu for thoughtful comments and assistance.

6. REFERENCES

[1] T. Barnes. The Q-matrix method: Mining student response data for knowledge. In *American Association for Artificial Intelligence Educational Data Mining Workshop*, 2005.

[2] M. Bernacki and S. Ritter. Motivation for learning HS Geometry 2012 (geo-pa). Dataset 748 in DataShop. Retrieved from `https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=748`, 2014.

[3] H. Cen, K. Koedinger, and B. Junker. Learning factors analysis – a general method for cognitive model evaluation and improvement. In *Proceedings of 8th International Conference on Intelligent Tutoring Systems*, pages 164–175, 2006.

[4] A. T. Corbett and J. R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4):253–278, 1995.

[5] A. Galyardt and I. Goldin. Recent-Performance Factors Analysis. In J. Stamper, Z. Pardos, M. Mavrikis, and B. McLaren, editors, *Proceedings of 7th International Conference on Educational Data Mining*, pages 411–412, 2014. (Poster paper).

[6] A. Galyardt and I. M. Goldin. Move your lamp post: Recent data reflects learner knowledge better than older data. *Journal of Educational Data Mining*, Accepted.

[7] R. Liu, K. Koedinger, and E. McLaughlin. Interpreting model discovery and testing generalization to a new dataset. In *Proceedings of 7th International Conference on Educational Data Mining*, 2014.

[8] R. C. Murray, S. Ritter, T. Nixon, and al. Revealing the learning in learning curves. In H. C. Lane, K. Yacef, J. Mostow, and P. Pavlik, editors, *Proceedings of 16th International Conference on Artificial Intelligence in Education*, pages 473–482, 2013.

[9] P. I. Pavlik, H. Cen, and K. Koedinger. Performance Factors Analysis–a new alternative to Knowledge Tracing. In *Proceedings of 14th International Conference on Artificial Intelligence in Education*, pages 531–538. IOS Press, 2009.

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

[11] K. K. Tatsuoka. Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20(4):345–354, 1983.