Toward Improving Student Model Estimates through Assistance Scores in Principle and in Practice

Napol Rachatasumrit
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
napol@cmu.edu

Kenneth R. Koedinger
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213
koedinger@cmu.edu

ABSTRACT

Student modeling is useful in educational research and technology development due to a capability to estimate latent student attributes. Widely used approaches, such as the Additive Factors Model (AFM), have shown satisfactory results, but they can only handle binary outcomes, which may yield potential information loss. In this work, we propose a new partial credit modeling approach, PC-AFM, to support multi-valued outcomes. We focus particularly on the amount of assistance, that is, the number of error feedback and hint messages, a student needs to get a problem step correct. Because errors and hint requests may not only derive from student ability, but also from non-cognitive factors (e.g., students may game the system), we first test PC-AFM on synthetic data where this source of variation is not present. We confirm that PC-AFM is indeed better than AFM in recovering the true student and knowledge component (KC) parameters and even predicts student error rates better than a model fit to error rates. We then apply the approach to six real-world datasets and find that PC-AFM outperforms AFM in reliable estimation of KC parameters and produces better generalization to new students, which requires better KC estimates. However, consistent with the hypothesis that student assistance behavior is driven by motivational or meta-cognitive factors beyond their ability, we found that PC-AFM was not better in reliable estimation of student parameters nor in generalization across items, which requires accurate student estimates. We propose *cross-measure cross-validation* as a general method for comparing alternative measurement models for the same desired latent outcome.

Keywords

Additive Factors Model, Student Modeling, Model Comparison, Learning Curves

Napol Rachatasumrit and Kenneth Koedinger "Toward Improving Student Model Estimates through Assistance Scores in Principle and in Practice". 2021. In: Proceedings of The 14th International Conference on Educational Data Mining (EDM21). International Educational Data Mining Society, 295-301. https://educationaldatamining.org/edm2021/
EDM '21 June 29 - July 02 2021, Paris, France

1. INTRODUCTION

Student modeling has been an important tool that researchers can use to estimate latent student abilities. Similarly, intelligent tutoring systems also depend on how accurately we can predict student mastery to deliver efficient adaptive learning. Current popular approaches, such as Additive Factors Model (AFM) [4, 18, 13] and Bayesian Knowledge Tracing (BKT) [5, 13], perform reasonably well by including the growth factors in their models. However, they are restricted by using only binary student performance (e.g. correct/incorrect response), which could suffer from an information loss due to its dichotomized nature.

For example, many existing intelligent tutoring systems (ITS) support step-by-step interactions [22], which usually allow students to try multiple attempts or request for hints until they are able to complete the step correctly. These interactions are important for an ITS because it allows the system to provide immediate feedback or support an adaptive experience, while collecting a rich interaction dataset on student actions. However, since AFM and BKT can only handle binary outcomes, the student data is needed to be aggregated through a rollup procedure before we can use it in student modeling. This means only success on students' first attempt on each step will be included in the data, and the rest of the actions (e.g. other attempt or hint requests) will be ignored. To illustrate how this could be problematic, let's imagine student A who had one incorrect attempt on a step before correctly completing it and student B who had multiple incorrect attempts and asked for multiple hints on the same step before getting it right. The dichotomous model like AFM and BKT would treat both students as the same on this particular step, but we can see that it is more likely that student A has demonstrated better knowledge than student B.

In our case, we are concerned with having a raw measure of student success at each assessment opportunity. There are different functions for producing or deriving an outcome measure for the data available in a tutoring system. Perhaps the most typical function is: first transaction correct = 1; otherwise = 0 where both hints and incorrect responses are both counted as a failure. While there are multiple ways to elicit polytomous outcomes from ITS student data, in this work we focus on an assistant score, which is a total number of incorrect attempts and hint requests combined for each step. From our preliminary analysis, we found that

there are correlations between assistance scores and AFM's predicted error rate, which suggests that there could be an extra information in assistance scores compared to a binary correctness outcome.

In this work, we are interested in whether or not an assistance score model could be a better predictor of student's change in performance than a dichotomous model like AFM. Particularly, our research questions are: (1) How can we develop an effective statistical measurement model that uses assistance scores? and (2) How do we compare two different response models?

A popular approach to compare different cognitive models in Educational Data Mining is to use goodness-of-fit (e.g. Bayesian Information Criterion), but it is not applicable in our scenario because our model is based on different outcomes (correctness vs assistance score). Alternative versions of measures of predictor variables can be contrasted through cross validations, but it becomes inadequate when the outcome variables are different. We also discuss a set of strategies for addressing the general problem of how to compare alternative measurement models for the same desired latent outcome. Particularly, how do we compare a binary correctness model with a polytomous Assistance Score model?

We propose a new cognitive modeling approach to support polytomous outcomes and demonstrated its ability to recover parameters and predict student error rates better than AFM in synthetic data. We then evaluated our model to six real-world datasets spanning five different domains from the DataShop repository [10]. We found that our model outperforms AFM in most Student-blocked CVs and estimating KC parameters, but it falls short at estimating student intercepts. We hypothesize that our model is struggling to estimate student parameters in the real-world datasets due to variance in students' help-seeking behavior, such as gaming-the-system, that leads to the extra variance in Assistance Scores above and beyond the variance associated with student ability.

2. RELATED WORK

2.1 Item Response Theory with Partial Credit

Item Response Theory (IRT) models [6] is the preferred method used in several state assessments in the United States and international assessments [8]. The goal of the IRT model is to estimate the latent construct (e.g. student ability) and item characteristics (item difficulty) based on only a collection of responses.

The simplest variation of IRT is the Rasch model (1PL model) [19], which is characterized by a single parameter representing item difficulty (d_j), and a single parameter representing student ability (a_i). As Eq.1 is equivalent to a logistic function, the Rasch model is essentially a logistic regression model.

$$p(r_{ij} = 1) = \frac{1}{1+^{-(a_i - d_j)}} \quad (1)$$

Other variations increase the complexity by introducing extra parameters. For example, the 2PL model adds a discrimination parameter for each item that controls the slope of the logistic function, and the 3PL model that also includes a pseudo-guessing parameter for each item. Even though, these models are characterized by a different number of parameters, they are all based on dichotomous response data (e.g. correctness). There is another class of IRT models that can be applied to polytomous outcomes, where each response can be a different value [17, 21]. An example of responses that is applicable to this class of models are Likert scale. There are different variations of polytomous IRT models, such as Partial Credit Model (PCM) [14], Generalized Partial Credit Model (GPCM) [15], and Graded Response Model (GRM) [20].These polytomous models are generalized from the dichotomous IRT models and can be reduced to the dichotomous IRT models when there are only two response categories. Our model extends the polytomous model to include growth factor by applying a similar approach to PCM to AFM.

2.2 Knowledge Tracing Approaches

Intelligent tutoring systems (ITS) have been shown to be effective in improving student learning outcomes across different domains [2, 9], and mastery learning strategies have been an important component in these systems. To implement mastery learning, knowledge tracing techniques are regularly utilized by ITSs [7] to adaptively assess students' knowledge states, which is used to decide when students have mastered skills and are ready to move on to other skills.

In many existing ITSs, such as Cognitive Tutor Authoring Tools (CTAT) [1], students are given a number of practice opportunities for each skill , and students are usually allowed to try multiple attempts or request for hints until they are able to successfully complete the step on each practice opportunity. The goal of a knowledge tracing algorithm when used for mastery learning is to determine when to stop giving students practice opportunities for the given skill.

Knowledge tracing is often performed by a statistical model of student learning that could be fit to data. There are two popular families of methods [12]: Bayesian Knowledge Tracing (BKT) [5, 13] and Additive Factors Model (AFM) [4, 18, 13]. Both methods include growth factors in order to estimate students' performance as it is changing with learning. BKT models student knowledge as a latent variable in a Hidden Markov Model. AFM is an extension of the IRT model that includes learning opportunity counts in the model. Even though these methods have been proven to work well in many scenarios, they are based on the binary error measurement model (correct or incorrect) and thus do not make use of potential added information from the number of error and hint messages a student may receive. Our approach explores this opportunity by extending AFM to use such multi-valued or polytomous outcomes in hopes of better estimating student knowledge. While other variations on AFM, such as Performance Factor Analysis (PFA) [18] and individualized AFM (iAFM) [13], have been shown in some cases to produce better prediction fit than AFM, we chose to use AFM to simplify the contrast between binary and polytomous measurement models and with the goal of producing more parsimonious and interpretable parameter estimates. Future work can explore alternatives.

2.3 DataShop Data Features

In this work, we use a variety of real world datasets across different domains from the DataShop repository [10]. LearnLab's DataShop (http://learnlab.org/datashop) is an open data repository of educational data with associated visualization and analysis tools, which has data from thousands of students derived from interactions with on-line course materials and intelligent tutoring systems.

In DataShop terminology, Knowledge Components (KCs) are used to represent pieces of knowledge, concepts or skills that students need to solve problems [11]. When a specific set of KCs are mapped to a set of instructional tasks (usually steps in problems) they form a KC Model, which is a specific kind of student model.

Each dataset in DataShop consists of a set of student transactions, which is a collection of students' interactions with ITSs. The collected students' actions include (but not limited to) correct attempts, incorrect attempts, and hint requests. The transactions that belong to the same practice opportunity get aggregated into a single students' step through the rollup procedure. The correctness of the step depends on the result of the student's first response for the practice opportunity, and the total number of incorrect attempts and hint requests is reported as an Assistance Score of the step. Most existing knowledge tracing algorithms use students' steps, rather than transactions, in their models.

3. METHOD

The Additive Factors Model (AFM) [4] is a logistic regression that extends Item Response Theory by incorporating a growth or learning term. The model 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 of the related KCs on the problem step (β_k), and the learning rate of the KCs (γ_k). The learning rate represents the improvement on a KC with each additional practice opportunity, so it is multiplied by the number of practice opportunities (T_{ik}) that the student already had on the KC.

$$log(\frac{p_{ij}}{1-p_{ij}}) = \theta_i + \Sigma_k(q_{jk}\beta_k + q_{jk}\gamma_k T_{ik}) \tag{2}$$

Our extension of AFM to support a polytomous outcome measure, like Assistance Score, is inspired by the Partial Credit Model (PCM) [14], which is an adjacent-categories logit model [21]. The model was designed to work with ordered polytomous response categories with a specific order or ranking of responses, which is the case for Assistance Score. It is widely applied in aptitude testing to allow for partial credit for near correctness of a response. In adjacent-categories logit models, we model the odds of a higher category relative to the adjacent lower one, and this paired comparison creates the ordering of the categories.

Assistance Score can be interpreted in the partial credit framework as follows. A student who gets a problem step correct on their first try or after fewer errors or hint requests is more likely to have the associated competence than a student who makes many errors or requests multiple hints before getting the step correct. Thus, students making no errors and needing no hints get full credit (Assistance Score = 0) and students with errors and/or hint requests get partial credit in rough proportion to the number hint and errors.

The Partial Credit Additive Factors Model (PC-AFM) builds upon these two different statistical models, AFM and PCM. For a student i and a step j, there is a set of probabilities $P_{ij} = \{p_{ija}; a = 0, 1, ..., A\}$ describing the chance for student i to get Assistance Score a on the step j, where A is the maximum Assistance Score. In this work, we decided to limit an Assistance Score at 5 because values above this tend not to be meaningful and rare, but extreme outliers (e.g., where assistance score is over 20 or even 140!) would significantly bias the model. 98% of our data have an Assistance Score of 5 or less. We extend AFM to use multivariate generalized linear mixed model, and the link function in logistic regression takes the vector-valued form.

$$f_{link}(P_{ij}) = \begin{pmatrix} f_{link,1}(P_{ij}) \\ ... \\ f_{link,A}(P_{ij}) \end{pmatrix} = \begin{pmatrix} log(\frac{p_{ij1}}{p_{ij0}}) \\ ... \\ log(\frac{p_{ijA}}{p_{ijA-1}}) \end{pmatrix} \tag{3}$$

Note that $f_{link,0}$ is not included due to the number of non-redundant probabilities. PC-AFM use adjacent-categories logits as a link function based on PCM. The ath adjacent-categories logit is the logit of getting an Assistance Score a versus $a-1$. Each link function is an extended version of AFM's linear model (Eq. 2) with a level parameter (α_a), which represents the difficulty to improve from an Assistance Score a to $a-1$.

$$f_{link,a}(P_{ij}) = \theta_i + \alpha_a + \Sigma_k(q_{jk}\beta_k + q_{jk}\gamma_k T_{ik}) \tag{4}$$

Inverting this function gives an expression for the probabilities of student i to complete a problem step j with each of the possible Assistance Scores a.

$$\begin{aligned} p_{ija} &= \frac{e^{\lambda_a}}{\Sigma_{i=0}^{A} e^{\lambda_i}} \\ \lambda_a &= \begin{cases} 0 & \text{if } a = 0 \\ \Sigma_{l=1}^{a} f_{link,l}(P_{ij}) & \text{otherwise} \end{cases} \end{aligned} \tag{5}$$

4. EXPERIMENT

We conduct experiments on both synthetic data and real student data to evaluate the performance of PC-AFM. We used the synthetic data to validate PC-AFM's parameter recovery capability and examine our evaluation strategy in a synthetic environment in which Assistance Score is stochastically derived from student ability alone. In particular, Assistance Scores in the synthetic data are not confounded by other student variations, such as their motivational state. We hypothesized that PC-AFM would work less effectively with the real student data because of non-ability effects on Assistance Score, such as students' help seeking strategies or propensity to game the system.

While goodness-of-fits metrics, such as BIC, are widely used

to compare different cognitive models [16], such as knowledge tracing algorithms, it is not applicable in our case due to the difference of outcome measures between AFM and PC-AFM. The challenge is how we can compare models that are based on different outcomes (error rate vs Assistance Score), while targeting the same desired latent measure (e.g. student's ability).

We explore two strategies to tackle this comparison problem. The first approach is to use parameter estimate reliability in split-half comparisons. Since both AFM and PC-AFM share the majority of their parameters (student intercepts, KC intercepts, and KC slopes), we can compare their parameter recovery capability. However, unlike synthetic data, the true parameters are not known in real data, so we need to use the reliability of parameter estimates in split-half comparisons instead. Another strategy is to compare cross-measure predictions. The assumption is that if a model based on polytomous outcomes (Assistance Score) yields better accuracy than a model based on binary outcomes (error rate) in predicting both polytomous and binary outcomes, the polytomous model will be demonstrated to be a better measurement model. This strategy is applicable in our scenario because there are connections between both outcomes. Since a student step is considered correct only when there is no assistance, the error rate can be derived by calculating the probability of Assistance Score = 0. On the other hand, we can convert the error rate to a probability of an Assistance Score by calculating the likelihood, where given an error rate p, the probability of having an Assistance Score a is $(1-p)p^a$. Then we can use CVs on both measures to compare the models.

4.1 Experiment 1: Synthetic Data

In order to validate PC-AFM capability to recover student and KC parameters, we synthetically generate datasets of student steps based on a logistic regression model. Given a set of student and KC parameters together with an opportunity count, a distribution over Assistance Scores is determined. We then sample once from the distribution to generate an Assistance Score of that student step. We generated 6 datasets of varying numbers of students and KCs, of which the true student and KC parameters are known, to examine parameter recovery capacity of PC-AFM in comparison to AFM. In each generated dataset, student intercepts range from -2 to 2, KC intercepts range from -1 to 1, and KC slopes range from 0 to 0.5. The number of KCs ranges from 8 to 32, and the number of students range from 25 to 200.

We also evaluate both models with three types of cross-validation (CV), Random (data points are split randomly), Student-blocked (data points are split by student), and Item-blocked (data points are split by item), to demonstrate if our model training on Assistance Score, can outperform a dichotomous model training on error rate in predicting dichotomous outcomes.

Table 1: Correlation between true and estimated parameters in synthetic data.

Dataset	Stu Intercept		KC Intercept		KC Slope	
	PC	AFM	PC	AFM	PC	AFM
KC8_S25	**0.978**	0.954	**0.996**	0.802	**0.914**	0.675
KC8_S50	**0.973**	0.936	**0.998**	0.985	**0.972**	0.964
KC8_S100	**0.973**	0.931	**1.000**	0.984	**0.952**	0.909
KC8_S200	**0.975**	0.936	**1.000**	0.979	**0.975**	0.735
KC16_S50	**0.990**	0.977	**0.998**	0.780	**0.962**	0.933
KC32_S50	**0.996**	0.988	**0.995**	0.799	**0.929**	0.543

Table 2: Correlation between split-halves parameters in synthetic data

Dataset	Stu Intercept		KC Intercept		KC Slope	
	PC	AFM	PC	AFM	PC	AFM
KC8_S25	**0.932**	0.828	**0.990**	0.895	**0.912**	0.498
KC8_S50	**0.963**	0.906	**0.998**	0.931	**0.972**	0.945
KC8_S100	**0.980**	0.941	**0.998**	0.850	**0.969**	0.888
KC8_S200	**0.871**	0.790	**0.999**	0.955	**0.910**	0.894
KC16_S50	**0.947**	0.857	**0.997**	0.947	**0.927**	0.843
KC32_S50	**0.967**	0.942	**1.000**	0.883	**0.997**	-0.345

We report on results for each of six different synthetic datasets by comparing PC-AFM and AFM. We found that PC-AFM better recovers the true student and KC parameters than AFM in almost all comparisons using correlation (Table 1). All contrasts are the same using mean absolute error. As the number of students goes up, both models tend to better recover the true parameters. The correlations of parameters in split-half comparison are reported in Table 2, which show a similar pattern to the correlation between estimated and true parameters. This demonstrates that the parameter correlation in split-half comparisons, which can be computed in real data, is a reasonable proxy for true parameter recovery, which cannot be computed in real data.

Figure 1 illustrates better true parameter recovery using Assistance Score and PC-AFM than using error rate and AFM. PC-AFM parameter estimates (red x's) are generally accurate across the spectrum of known parameter values (x-axis), as can be seen by their closeness to the line, which is identity function (intercept of 0, slope of 1). AFM estimates (blue dots) are generally biased toward the extremes. For student intercepts (Figure 1a), low prior knowledge students are estimated by error rate/AFM to be worse than they are and high prior knowledge students are estimated to be better than they are. For KC intercepts (Figure 1b), hard KCs (on the left) are estimated by error rate/AFM to be even harder than are. For hard KCs, most responses are errors, yielding quite low estimates by error rate/AFM. But, these same steps show more variance in Assistance Score/PC-AFM as somewhat better students and higher opportunities will produce lower, but non-zero Assistance Scores (i.e., not changing in error rate).

In error rate CV results, except Item-blocked CV where both models perform similarly, PC-AFM outperforms AFM in all other CVs (Table 4). Recall that these CV evaluations require PC-AFM, while fit to Assistance Score (polytomous outcome), to predict error rate (dichotomous outcome). When we turn the tables and compare methods on predicting Assistance Score, we find a similar pattern where PC-AFM yields better accuracy in most CVs (Table 3).

4.2 Experiment 2: Real student data

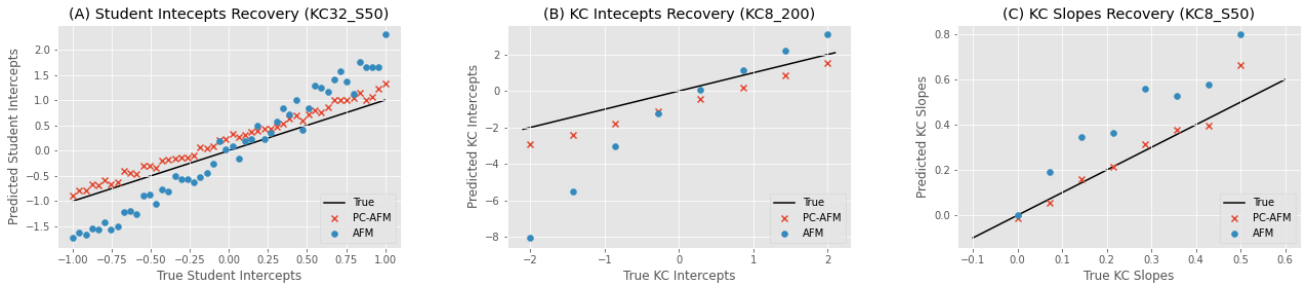

Figure 1: Using Assistance Score and PC-AFM on synthetic data produces better estimates of the true parameters, for all three of student intercepts, KC intercepts, and KC slopes than does using error rate and AFM.

Table 3: Cross-validation results (RSME) in synthetic data predicting Assistance Score in the test set by estimating parameters based on Assistance Score (PC-AFM) or on Error Rate (AFM) in the training set.

Dataset	Random		Stu-Blocked		Item-Blocked	
	PC	AFM	PC	AFM	PC	AFM
KC8_S25	**0.546**	0.598	**0.542**	0.600	**0.586**	0.634
KC8_S50	**0.544**	0.599	**0.541**	0.601	**0.575**	0.610
KC8_S100	**0.536**	0.596	**0.532**	0.599	**0.550**	0.602
KC8_S200	**0.541**	0.597	**0.537**	0.600	**0.541**	0.597
KC16_S50	**0.540**	0.600	**0.537**	0.601	**0.566**	0.604
KC32_S50	**0.540**	0.587	**0.539**	0.590	**0.579**	0.626

In the second experiment, we examine PC-AFM across a variety of real world datasets. We used 6 datasets across different domains (statistics, English articles, algebra, and geometry) from the DataShop repository. Table 5 shows the number of students, items, KCs, total transactions for each dataset. For each dataset, we use the KC model that achieves the best BIC reported on the DataShop repository. All KC models coded a single KC per step. The number of KCs ranges from 9 to 64, and the number of students ranges from 52 to 318.

For each dataset, we evaluated both PC-AFM and AFM on 5 independent runs of 3-fold CVs of each type predicting both Assistance Score and error rate. We report the result of Assistance Score CVs in Table 6 and the results of error rate CVs in Table 7. We found that PC-AFM outperforms AFM in Student-blocked in both Assistance Score and error

Table 4: Cross-validation results (RSME) in synthetic data predicting Error Rate in the test set by estimating parameters based on Assistance Score (PC-AFM) or on Error Rate (AFM) in the training set.

Dataset	Random		Stu-Blocked		Item-Blocked	
	PC	AFM	PC	AFM	PC	AFM
KC8_S25	**0.275**	0.278	0.310	**0.306**	**0.370**	0.430
KC8_S50	**0.273**	0.280	**0.282**	0.304	0.356	**0.297**
KC8_S100	**0.273**	0.277	**0.283**	0.300	**0.387**	0.449
KC8_S200	**0.271**	0.275	**0.278**	0.295	**0.278**	0.282
KC16_S50	**0.277**	0.281	**0.278**	0.311	0.301	**0.294**
KC32_S50	**0.287**	0.291	**0.292**	0.320	0.358	**0.347**

Table 5: Real Student Dataset.

Dataset	Domain	#Stu	#Item	#KC
ds308	College Statistics	52	113	9
ds313	English articles	120	85	26
ds372	English articles	99	84	15
ds388	Middle School math	318	64	64
ds392	Geometry	123	2035	43
ds394	English articles	97	180	13

rate CVs in most datasets, which suggests that PC-AFM can achieve better estimates of KC parameters. To validate the hypothesis, we investigated split-halves parameters correlation of both models. We splitted the datasets on students to evaluate KC slopes and intercepts correlation, and we splitted the datasets on KCs to evaluate students' intercepts (Table 8). On average, PC-AFM yields better correlations of both KC intercepts (0.954 vs 0.946) and KC slopes (0.600 vs 0.563), but correlations of student intercepts is significantly higher for AFM (0.784 vs 0.495).

5. DISCUSSION

Assistance score should, in principle, improve model parameter estimates and predictions based on them. A student who gets a step correct after just one error or one hint (Assistance Score = 1) is likely to be closer to full acquisition of a KC than a student who makes an error and requests 3 hints (Assistance Score = 4). However, the error rate metric commonly used with BKT and AFM treats these the same, since the student was not correct on their first attempt at the step without a hint. Thus, there is potentially extra in-

Table 6: Cross-validation results (RSME) in real data predicting Assistance Score in the test set by estimating parameters based on Assistance Score (PC-AFM) or on Error Rate (AFM) in the training set.

Dataset	Random		Stu-Blocked		Item-Blocked	
	PC	AFM	PC	AFM	PC	AFM
ds308	0.376	0.376	0.381	**0.378**	**0.384**	0.388
ds313	0.541	**0.528**	**0.551**	0.554	**0.549**	0.555
ds372	0.478	**0.463**	**0.480**	0.481	**0.484**	0.487
ds388	0.672	**0.649**	**0.682**	0.703	**0.702**	0.703
ds392	0.385	**0.354**	**0.386**	0.387	**0.385**	0.390
ds394	0.499	**0.486**	0.499	0.499	**0.504**	0.510

Table 7: Cross-validation results (RSME) in real data predicting Error Rate in the test set by estimating parameters based on Assistance Score (PC-AFM) or on Error Rate (AFM) in the training set.

Dataset	Random		Stu-Blocked		Item-Blocked	
	PC	AFM	PC	AFM	PC	AFM
ds308	0.336	**0.326**	0.332	**0.328**	0.341	**0.339**
ds313	0.417	**0.408**	**0.413**	0.440	0.435	**0.424**
ds372	0.379	**0.377**	**0.383**	0.402	0.388	**0.387**
ds388	0.454	**0.421**	0.439	**0.470**	0.501	**0.456**
ds392	0.324	0.324	**0.325**	0.333	0.325	0.325
ds394	0.395	**0.391**	**0.388**	0.418	0.403	0.403

formation about students' level of knowledge acquisition in the Assistance Score not present in error rate. On the other hand, prior research, for example on gaming the system [3], suggests there are other reasons students may produce repeated incorrect entries or hint requests. These may produce enough confounding variance to make using Assistance Score worse at accurate latent parameter estimation than using error rate.

In developing a statistical model, PC-AFM, to convert Assistance Scores to knowledge acquisition estimates, we first wanted to confirm that PC-AFM works as intended and is able to benefit from extra information in Assistance Score when no confounding sources for Assistance Score variation are present. Indeed, when we generate synthetic data where Assistance Scores are stochastically produced from known latent parameters, we demonstrate better parameter recovery using Assistance Score and PC-AFM than using error rate and AFM. As shown in Figure 1, PC-AFM estimates of student parameters are better correlated with true parameters and the AFM estimates are baised at the extremes.

This parameter recovery method for comparing these two different measurement models cannot be applied to real datasets because the true parameters are unknown. Thus, we employed we explored two other approaches: parameter estimate reliability and our novel cross-measure cross-validation approach. We demonstrated better parameter estimate reliability (in split-halves comparisons) using PC-AFM than AFM. We also show how it is possible to use cross-measure predictions to evaluate which of two different measurement models works better, call them M1 and M2. We show that estimating based on M1 (e.g., assistant score) can predict M2 (e.g., error rate) on held-out data better than estimating based on M2 itself (e.g., error rate). We believe this cross-measure cross-validation is a novel approach for comparing measurement models.

Assessing whether Assistance Score is a better measure than Error Rate in real student data is complicated in two ways. First, we do not have access to the true parameters in real datasets, so we turn to measures of reliability and predictive validity. Second, we know from models of gaming the system and help seeking that students may produce Assistance Scores for motivational and metacognitive reasons that are potentially independent of a mastery source. In other words, Assistance Scores have a student-driven source of variation that may reduce their effectiveness in estimating student mastery. We hypothesize that our model is struggling to estimate student parameters in the real-world datasets due to variance in students' help seeking behavior.

Table 8: Split-halves parameters correlation in real data.

Dataset	Stu Intercept		KC Intercept		KC Slope	
	PC	AFM	PC	AFM	PC	AFM
ds308	0.113	**0.486**	**0.971**	0.955	**0.745**	0.583
ds313	0.490	**0.830**	**0.948**	0.937	0.865	**0.905**
ds372	0.427	**0.803**	**0.985**	0.968	0.433	**0.639**
ds388	0.567	**0.873**	**0.946**	0.945	0.225	**0.354**
ds392	0.830	**0.901**	**0.973**	0.964	**0.494**	0.485
ds394	0.541	**0.809**	0.904	**0.906**	**0.838**	0.413

We found that in real world datasets PC-AFM can better estimate KC parameters than AFM, which results in PC-AFM outperforming AFM in Student-blocked CVs. KC parameters estimates significantly impact Student-blocked CVs because they are the sole driver of these predictions. Poor student estimates do not impact Student-blocked CVs because they are not carried from the training to test as blocking means there are different students in the test than training. It does impact Random CVs and Item-blocked CVs because they are likely to have some students showing up in both test and training.

6. CONCLUSION AND FUTURE WORK

We investigated whether or not Assistance Score provides a better measurement model than error rate for estimating student's ability. To pursue this question, we developed a statistical model, PC-AFM, that utilizes Assistance Score. We also faced the more general problem of how to compare alternative measurement models for the same desired latent outcome. In typical model comparison the predicted outcome measure stays the same, but such comparison does not work when the outcome measures are different. We proposed two strategies to tackle this problem: parameter estimate reliability in split-halves comparisons and a new approach we call, cross-measure cross-validation. We demonstrated that these strategies work well by using synthetic data to show that a model that better recovers parameters will also yield better results with these strategies.

We demonstrated that PC-AFM outperforms AFM when Assistance Scores are synthesized to be meaningful, but its performance is hindered by non-ability variance in students' behavior in the real-world datasets. Future work can explore this finding by synthesizing Assistance Scores that derive from both ability and motivational factors.

Future work can also test our measurement model comparison strategies. For example, while it has been standard practice in many tutoring systems to count hints as errors (M1), some have wondered whether it would be better to not count hints as errors (M2). Our measurement model comparison techniques, split-half reliability and cross-measure cross-validation, can be used to compare M1 and M2 to infer which provides better estimates of student ability.

7. REFERENCES

[1] V. Aleven, B. M. McLaren, J. Sewall, and K. R. Koedinger. The cognitive tutor authoring tools (ctat): Preliminary evaluation of efficiency gains. In *International Conference on Intelligent Tutoring Systems*, pages 61–70. Springer, 2006.

[2] J. R. Anderson, A. T. Corbett, K. R. Koedinger, and R. Pelletier. Cognitive tutors: Lessons learned. *The journal of the learning sciences*, 4(2):167–207, 1995.

[3] R. Baker, J. Walonoski, N. Heffernan, I. Roll, A. Corbett, and K. Koedinger. Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, 19(2):185–224, 2008.

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

[5] R. S. d Baker, A. T. Corbett, and V. Aleven. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In *International conference on intelligent tutoring systems*, pages 406–415. Springer, 2008.

[6] S. E. Embretson and S. P. Reise. *Item response theory*. Psychology Press, 2013.

[7] Y. Gong and J. E. Beck. Towards detecting wheel-spinning: Future failure in mastery learning. In *Proceedings of the second (2015) ACM conference on learning@ scale*, pages 67–74, 2015.

[8] W. Harlen. The assessment of scientific literacy in the oecd/pisa project. 2001.

[9] K. R. Koedinger, J. R. Anderson, W. H. Hadley, M. A. Mark, et al. Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*, 8(1):30–43, 1997.

[10] K. R. Koedinger, R. S. Baker, K. Cunningham, A. Skogsholm, B. Leber, and J. Stamper. A data repository for the edm community: The pslc datashop. *Handbook of educational data mining*, 43:43–56, 2010.

[11] K. R. Koedinger, A. T. Corbett, and C. Perfetti. The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive science*, 36(5):757–798, 2012.

[12] K. R. Koedinger, S. D'Mello, E. A. McLaughlin, Z. A. Pardos, and C. P. Rose. Data mining and education. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(4):333–353, 2015.

[13] R. Liu and K. R. Koedinger. Towards reliable and valid measurement of individualized student parameters. *International Educational Data Mining Society*, 2017.

[14] G. N. Masters. A rasch model for partial credit scoring. *Psychometrika*, 47(2):149–174, 1982.

[15] E. Muraki. A generalized partial credit model: Application of an em algorithm. *ETS Research Report Series*, 1992(1):i–30, 1992.

[16] A. A. Neath and J. E. Cavanaugh. The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203, 2012.

[17] R. Ostini and M. L. Nering. *Polytomous item response theory models*. Number 144. Sage, 2006.

[18] P. I. Pavlik Jr, H. Cen, and K. R. Koedinger. Performance factors analysis–a new alternative to knowledge tracing. *Online Submission*, 2009.

[19] G. Rasch. Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests. 1960.

[20] F. Samejima. Graded response model. In *Handbook of modern item response theory*, pages 85–100. Springer, 1997.

[21] F. Tuerlinckx and W.-C. Wang. Models for polytomous data. In *Explanatory Item Response Models*, pages 75–109. Springer, 2004.

[22] K. VanLehn. The behavior of tutoring systems. *International journal of artificial intelligence in education*, 16(3):227–265, 2006.